

Meredith L. Hale. Searching for Art Records: A Log Analysis of the Ackland Art Museum's Collection Search System. A Master's Paper for the M.S. in I.S. degree. August, 2015. 88 pages. Advisor: Diane Kelly.

Search log data from the Ackland Art Museum's online collection search system was analyzed in order to determine the search categories most frequently employed by users of the system. The data consisted of a total of 16,729 actions and 3,459 search sessions. It covered a three-month time period from February 19 to May 19, 2015. Analysis of actions associated with the Ackland's advanced search feature suggest that the department, classification, and artist fields have the highest usage while searchers rarely submit queries relating to a work's particular medium (2.55%). Review of the most common queries submitted by users reveals that search terms most commonly relate to representational subjects visually presented in a work of art rather than formal titles. Investigation into how users alter the queries they submit throughout a search session indicates that users often do not change categories during a search session, but primarily make parallel changes (68.11%).

Headings:

Art museum websites -- Use and access

Server-side user study -- Art museum online collection search system

Information needs -- Art historians, artists, and humanities scholars

Query reformulation

SEARCHING FOR ART RECORDS:
A LOG ANALYSIS OF THE ACKLAND ART MUSEUM'S COLLECTION SEARCH
SYSTEM

by
Meredith L. Hale

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

August 2015

Approved by

Diane Kelly

TABLE OF CONTENTS

Acknowledgements.....	2
List of Figures & Tables	3
Introduction.....	4
Overview of Art Museum Collection Search Systems	7
Description of the Ackland Art Museum's Search System & Collection	14
Literature Review.....	22
Methods.....	34
Method Selection: Log Analysis and its Strengths & Weaknesses	34
Documentation of Decisions for Data Cleaning & Analysis	36
Results & Discussion	42
General Characteristics of the Log Data.....	42
Analysis of Advanced Search Category Usage	49
Corpus of Queries	54
Query Reformulation	64
Conclusions.....	69
Recommendations.....	71
Limitations	76
Postscript.....	78
Bibliography	80
Art Captions	85

ACKNOWLEDGEMENTS

Thanks to:

Dr. Diane Kelly for teaching me what a log analysis is, instigating my interest in information retrieval, and making me see how valuable user-centered research can be

Joan Boone for introducing me to programming and helping me apply my knowledge to the world of art

Patrick Golden for making me realize how messy data can be and suggesting tools to clean it up

Scott Hankins and the staff at the Ackland Art Museum for providing me with the data that made this study possible

LISTS OF FIGURES & TABLES

List of Figures

- Figure 1 - Example collection highlights page from the Dayton Art Institute (DAI)
- Figure 2 - Detroit Institute of Art's collection search system
- Figure 3 - The Indianapolis Museum of Art's (IMA) collection search page
- Figure 4 - Dallas Museum of Art's (DMA) collection search page
- Figure 5 - San Francisco Museum of Modern Art (SFMOMA) ArtScope search feature
- Figure 6 - Layout of the Ackland's collection search system
- Figure 7 - Detailed view of the Ackland's advanced search feature
- Figure 8 - Ackland's collection highlights page
- Figure 9 - Example action from the log submitted by a robot to illustrate its structure
- Figure 10 - Timeline of the Ackland's search log created using OpenRefine
- Figure 11 - Percentage of search sessions by action total
- Figure 12 - Distribution of search sessions with less than 50 actions
- Figure 13 - Percentage of actions associated with the three search types
- Figure 14 - Percentage of advanced search categories associated with terms found within unique queries by session
- Figure 15 - Maximum number of advanced search categories per search session by percentage
- Figure 16 - Word cloud of the top 49 search terms unique to a search session
- Figure 17 - Richard Westall, *The Sword of Damocles*, oil on canvas, 1812, Ackland Art Museum.
- Figure 18 - Categorical query reformulation types by percentage
- Figure 19 - Works in the Ackland's collection with the highest number of search terms related to them

List of Tables

- Table 1 - Duration and additional characteristics of search sessions with over 100 search actions
- Table 2 - Queries associated with the most actions
- Table 3 - List of the top queries by frequency that are unique in a search session

INTRODUCTION

In the development of a search system, detailed attention is frequently given to understanding and representing the “documents” a system will need to retrieve, but less consideration is often accorded to determining how the needs of potential users of a system will match with this document-focused metadata. Ultimately having a proper understanding of both users and the items they seek to retrieve is essential to creating an effective search system. The writings of Nicholas Belkin support this, especially emphasizing the importance of the user in system development. He writes, “the corpus of needs is at least as important as the corpus of documents in producing rules for representation and perhaps more so” (Belkin, 1980, p. 136). Request- or user-oriented approaches to indexing have also been forwarded by scholars such as Dagobert Soergel and Raya Fidel and have become increasingly prominent in collection management and information retrieval today (Fidel, 1994, p. 574).

Following in this vein, this study aims to increase understanding of the needs of users of the Ackland Art Museum’s collection search system. The main research question is to determine the types of search categories (artist name, date, etc.) most frequently submitted by users. In order to achieve this, log data from the Ackland Art Museum was collected and evaluated to uncover patterns in the type of content typically sought and the search methods frequently used by visitors to the site. The search URLs found within the log contain the query terms submitted by users as well as the search category they are

associated with in some instances, which can be used to indicate the types of information users are seeking. While these queries are not equivalent to an information need, they are indicative of these needs. The queries can be best understood as examples of Robert Taylor's definition of a compromised need (Taylor, 1968, p. 182). When a user must alter his or her actual information need in order to make it acceptable to the structure of a particular system, this need is defined as compromised. Almost any need that is recorded in some way is necessarily compromised because of the "translation" that occurs in formalizing a visceral need. Yet, while this analysis documents compromised needs, they are *real* needs. The use of log data in this study makes this authenticity possible. Using log data, rather than interviewing users about their needs retrospectively or using task-based scenarios to elicit genuine information-seeking behaviors, has value because of its ability to unobtrusively document the actual queries of current users of an existing search system. Further discussion of the strengths and weaknesses of log analysis as a research method is included in the "Methods" section.

While uncovering user needs is always challenging, studying these needs in an online art museum collection has its own particular set of trials to be overcome due to its emphasis on visual information retrieval (VIR). While a visitor to an art museum's online collection search system may be interested in searching the system to gather either visual or textual material, both the text found in an art historical record and the visual representation of the object are tied to a visual rather than a textual source. Central to any study concerned with image retrieval is the largely unexplained phenomenon of how people perceive and verbalize their need for visual information. This process still remains unexplained simply because language cannot act as an exact surrogate for a non-verbal

entity. This failure of language is the cause of the primary challenge faced by visual information retrieval researchers today – the semantic gap. This term refers to the lack of coherence between an item’s visual characteristics and the interpretation of these characteristics by a user (Smeulders, Worring, Santini, Gupta, & Jain, 2000, p. 1353). Unlike text retrieval, “visual materials have broader and less well-defined access points because they can be described by a variety of factors within them” (Choi & Rasmussen, 2003, p. 498). Search systems, like the one used by the Ackland, must strive to predict how users will perceive and interpret images in order for retrieval to be effective.

While this study primarily aims to help close the semantic gap and better understand user needs, it is impossible to fully understand these needs without also having knowledge about the search system. In search, a balancing act is constantly taking place between users and systems. Especially in systems that do not permit natural language queries, users must alter how they would phrase their need in order to make it match existing search parameters. In theory, system designers are also always analyzing how users interact with their systems in order to find ways to make these systems more effective. In order to address this relationship, an overview of some of the current trends in the design and structure of art museum collection search systems follows. The Ackland’s collection search system will then be described and situated in relation to these broader general trends. Because the structure of a search system has the potential to greatly influence the behaviors of users and dictate the type of information for which users can search, this information is essential to have as a foundation before delving into the specific findings.

Overview of Art Museum Collection Search Systems

Art museums typically focus their gallery spaces narrowly on collection strengths and duplicate this practice on their websites through the widespread use of “collection highlights” features (Fig. 1). Often when multiple highlights pages are made for a single collection, cultural categorizations are used to organize the materials, despite the ideological controversy this can cause. While highlights pages arguably focus the attention of web visitors on the highest quality works, they are also often used without any additional features as a replacement for a more interactive and comprehensive search system. Highlights pages typically feature a handful of works from the collection and, because they do not usually include links or search boxes, do not encourage users to explore further. According to a study by Peacock, Ellis, and Doolan in which 100 museums were surveyed, 53 percent of museums only have static collection highlights pages (2004, p. 14). Several state universities, such as San Jose and Colorado, rely upon proprietary systems like EmbARK Web Kiosk to go beyond the limited access offered by collections highlight pages, but these systems also have drawbacks and aesthetically appear dated.¹ Still, more robust museum search systems do exist, especially in large institutions that receive government funding.

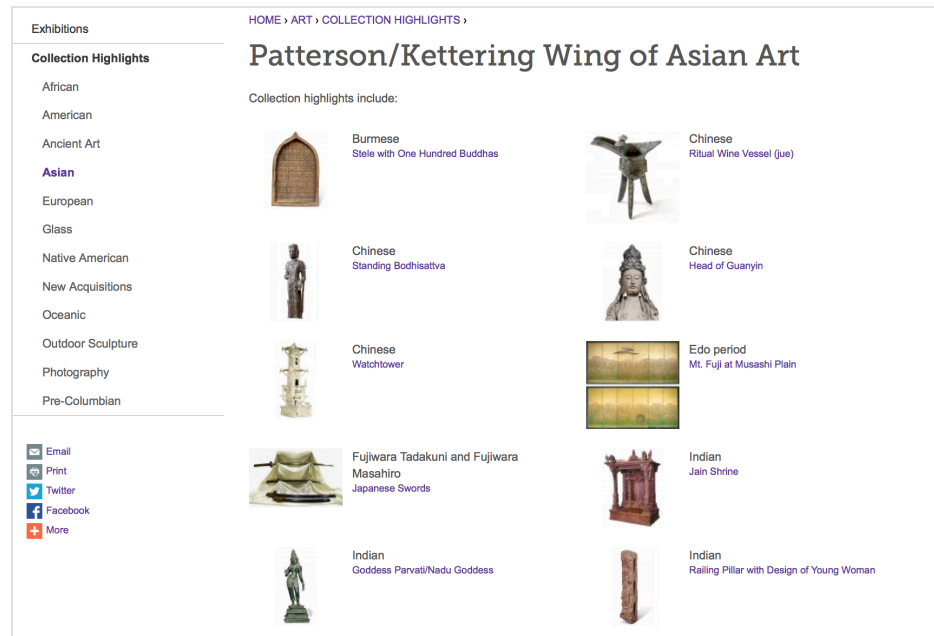


Figure 1 – Example collection highlights page from the Dayton Art Institute (DAI)

More contemporary museum websites often support user-inputted queries as well as browsing through the use of facets or images. While some museums still have complex advanced search options with many fields for user queries (Fig. 2), this kind of feature is not common. Instead the use of a single query box in conjunction with other features is more typical. The Indianapolis Museum of Art's home search page for their collection (Fig. 3), is an extreme example of how many institutions are replacing multiple search input fields with a single search box similar to search engines like Bing and Google. Further influence of commercial websites can also be seen in the growing prevalence of faceted search features. Categories frequently used for facets include date, culture, artist/maker, and object type/material. In addition, many museums have department and classification categories that reflect divisions that are standard in The Museum System (TMS) database. The Yale University Art Gallery, as well as the Ackland, have these categories.

Find and view your favorite works online

Note: Only a portion of the DIA's complete collection is represented online.

Select a Collection

Select a Classification

Select an Artist

Select an Artist Nationality

Select a Medium

ENTER AN ACCESSION NUMBER

ENTER ARTIST, TITLE, OR KEYWORD

SPECIFY A TIME PERIOD (4 DIGIT YEAR)
FROM TO

☐ SHOW ITEMS WITH OR WITHOUT PHOTOGRAPHS

[Search Art at the DIA](#) [Reset Form](#)




Figure 2 – Detroit Institute of Art’s collection search system, which features categories like accession number that are similar to those found on the Ackland’s system

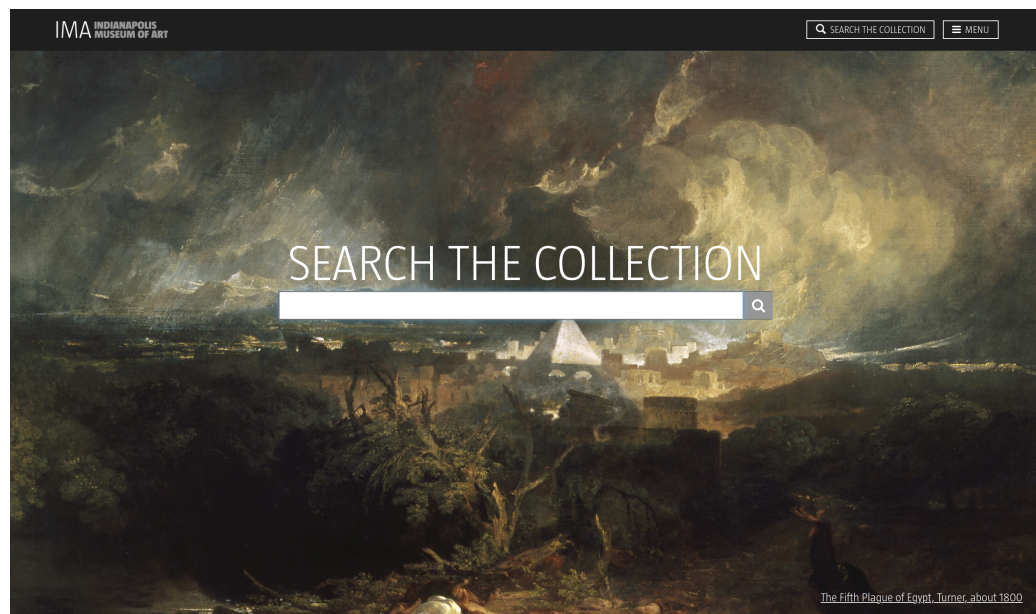


Figure 3 - The Indianapolis Museum of Art’s (IMA) collection search page
<http://collection.imamuseum.org/>

In addition to the standard art historical categories emphasized on both art museum websites and established thesauri like the Art and Architecture thesaurus (AAT), subject description is seen as a possible way to enrich access to art materials (Lunin, 1994, p. 67). According to Jørgensen, “abstract concepts, emotions, stories, and ‘people-related’ information such as social status would be useful in image retrieval” (Jørgensen, 1999, p. 348). Search systems that feature subject searching, while rare, are now offered by a few museums to provide users with ways to access the collection that are not specific to the discipline of art history. The best example of this is the Tate Museum and its subject thesaurus which includes broad categories like “emotions and human qualities” that allow the user to search for works that display ideas like “frustration,” “shyness,” or “wisdom.”² Including access points like these demonstrates a desire to make collections more open to those without specific art or art historical knowledge.

Other museums that provide subject access often rely on their users to generate terms in a kind of folksonomy. Among the most successful of these crowdsourcing efforts is the Walters Art Museum “tags” page.³ While some folksonomies have struggled with getting enough participation, the Walters Art Museum has successfully gathered thousands of tags from users. Submissions range from art historical terms like “mannerism” to more abstract feelings like “love.” Misspellings are common, for instance there are five different tags of the word “awesome,” but it may be worth wading through some contributions of limited value for the unique access points this method offers.

In addition to starting to incorporate subject access, museum collection search systems are relying less on text and featuring more visual materials for discovery and

navigation. One very specific example of a visual development just beginning to be implemented on art museum collection search systems is the use of Content-Based Image Retrieval (CBIR). CBIR allows users to search for images within the collection by visual features, such as color, shape, and texture. CBIR is an alternative to textual keyword retrieval. In addition to the unique access points this technology creates, its ability to automate indexing through using similarity matching makes CBIR a compelling retrieval option. Yet, the visual features it focuses on are considered to have low semantic meaning when compared to text (Hare, Lewis, Enser, & Sandom, 2006, p. 2). This type of retrieval is relatively new, having been first implemented in 1994 with Query by Image Content (QBIC) (Liu, Zhang, Lu, & Ma, 2007, p. 262). Examples of the few museums that have incorporated this technology into their search systems include the Dallas Museum of Art (DMA), the Victoria and Albert Museum (V & A), the Rijksmuseum, the Indianapolis Museum of Art (IMA), and the Cooper-Hewitt Museum (Fig. 4).⁴ The Cooper-Hewitt Museum is the only system that currently supports searches for CBIR features beyond color. The State Hermitage Museum of Russia piloted IBM's QBIC in 2004 on their collection website, which allowed users to search by form in addition to color, but this feature is longer available on the museum's website (Wells-Angerer, 2005, p. 20). While CBIR offers unique access points into collection records, its use and effectiveness in an art museum context still remains largely unevaluated.

The Dallas Museum of Art's collection contains over 23,000 works of art from all cultures and time periods spanning 5,000 years of human creativity. The collection is dynamic; new

acquisitions are being added all the time and the galleries are constantly changing. **Notice and Disclaimer**

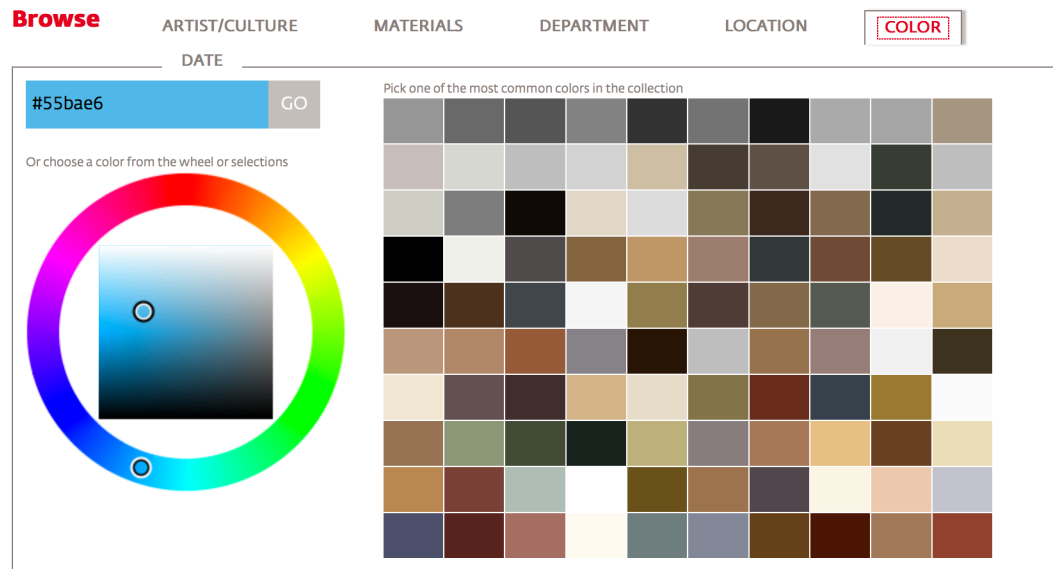


Figure 4 - Dallas Museum of Art's (DMA) collection search page featuring CBIR color search technology - <https://www.dma.org/collection>

In addition, images have become increasingly important to navigating online collection search systems. Mosaic pages, like those found on Pinterest and Tumblr, seem to have also influenced museum collection search pages. The San Francisco Museum of Modern Art (SFMOMA) recently developed ArtScope, which allows users to visually browse a wall of continuous images that represent a selection of the museum's larger collection (Fig. 5). Another feature that aims to increase the interaction between website visitors and museum objects through visual means is a random generator that showcases various art objects without taking into account any specific user need besides variety. These generators attempt to provide website visitors with a sense of serendipity as they navigate through the collection search system. All of these search features that focus on images should be well suited for users that seek information visually, but they do not

offer many alternatives for a diverse population of users if offered as the sole means of navigating the collection.

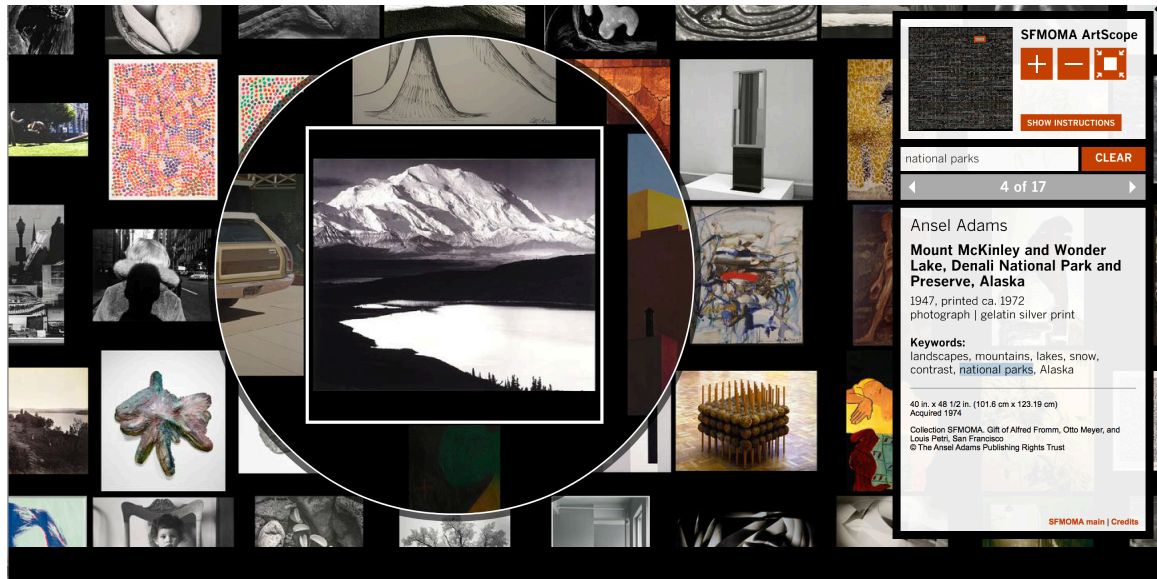


Figure 5 – San Francisco Museum of Modern Art (SFMOMA) ArtScope search feature - <http://www.sfmoma.org/projects/artscope/>

Finally, several museums have implemented new facets or search categories that, while less broad in their scope than subject searching or CBIR, are also significant. For instance, the Los Angeles County Museum of Art (LACMA) includes facets to let users search by the work's location within the museum. The Yale Center for British art also has a search category particularly focused on frames that allows users to look for a frame's maker or ornamental types included in the work.⁵

Whether CBIR, subject thesauri, or specific search parameters like accession number or medium should be included on online art museum collection search systems is ultimately dependent upon a system's users. If interviews with museum users or log data collected through interaction with a particular system reveal that online visitors want to be able to search the collection for abstract concepts like "harmony," the system should

be able to support these needs. Ideally, in order to both accurately document the collection and meet user needs, records on museum search systems should support a range of meanings for a single work that in turn provide users with multiple access points (Cameron, 2012, p. 228). Yet this must also be tempered with the reality that creating access points, especially through keyword indexing, is an extremely time-consuming process. Research on which access points are utilized most are therefore critical to using staff time effectively while achieving the ultimate goal of access. This research also serves the more theoretical goal of augmenting current knowledge on how people perceive art.

Description of the Ackland Art Museum's Search System & Collection

While the primary focus of this study is the user rather than system, the search system people interact with has the potential to greatly influence their search behaviors and dictate the types of information for which they can search. This, in turn, also affects what information can be gathered to study and which methods would be appropriate to analyze this data. It was decided to particularly focus on how users search for art objects on the Ackland Art Museum's online collection search system due to the museum's willingness to share their search log data with the primary investigator. That a study had never been conducted previously using the Ackland's search log also supported this decision. In order to address how the Ackland Art Museum's search system may affect the methods and results of this study, a brief overview of the system's history, a description of its current structure and capabilities, and an overview of its collection follow.

The Ackland Art Museum has had three different search interfaces in its history. That a smaller university-affiliated museum like the Ackland has a search system instead of only a static collection highlights page is commendable. The museum developed its website around 1999 and in the year 2000 the museum purchased its first collection management system, Multi MIMSY (Beard, 2004, pp. 1, 11). In 2004 a UNC Information Science student, Carmen Beard, created the first online collection search system as a Master's project (Beard, 2004, p. 1). In response to a staff survey conducted by Beard, the interface was made with the intention of supporting known-item searching over browsing (Beard, p. 43). This can especially be seen in the advanced search feature that will be described later in this section. The interface's exclusively textual retrieval system also continues to encourage known-item retrieval over exploratory search. A second search system had to be created in 2008 after the museum switched their database from MIMSY to The Museum System (TMS). An internal staff member created this system. An outside web developer was later hired to develop the Ackland's current search system during the 2010-2011 school year (Scott Hankins (Registrar at the Ackland Art Museum), personal communication, June 5, 2015). This "new" system built upon, but also improved, the existing search system.

The Ackland Art Museum's website supports three different ways for users to submit queries to retrieve records from the collection. While these three methods all rely upon the same PHP code, they do have notably distinct visual and functional features that have the potential to influence user searches. The museum's collection search page has two main search features: a simple (broad) search with a single-search box and an advanced (focused) search with a total of ten categories for user input (Fig. 6). While text

on the website describes these two features as broad and focused, I will be using the alternative terms “simple” and “advanced,” which reflects the terms used in the web log to describe these types of searches, when referencing these features for this study. The collection can additionally be queried from any page on the Ackland’s website by typing into a single-search box found in the upper-right-hand corner of each page and selecting a radio button named “Collection.” This search type will be referred to as “simple radio” throughout this paper because it runs the same basic keyword search algorithm as the simple search but differs in its presentation.

ACKLAND ART MUSEUM

Simple radio ☐ ☐ Ackland.org

Home About Visit On View Collections Education Calendar Support Contact Shop

Home » Collection » Collection Database Search

Welcome to the Ackland's Collection Database

From here you can search our database of records for every object in the permanent collection, including those on long-term loan to the Museum. You have two ways to search: a broad search by entering one or more keywords in the field below, and a more focused search using one or more of the categories in the form to the right.

Search:

Simple search

The Broad Search will provide a list of records which contain the keywords in any of a range of fields. For example, searching for "wood" will bring up a list of objects that includes works by an artist such as Grant Wood, wood engravings, objects made of wood, and works with the word "wood" in the title. Only complete instances of the word are selected.

The Focused Search allows you to specify more detail and to control the fields searched. Some fields offer a drop-down menu of choices (Department and Classification); others will automatically suggest possible options as you type (Artist, Title, Medium, and Culture). In Focused Search, parts of words are also considered. For example, searching on "Frank" in Title will return "Frankie and Johnny" and "Rotation Frankfurt III", as well as "Frank Kenan" and "Frank Porter."

Department: (any)
 Classification: (any)
 Artist/Creator:
 Title:
 Medium:
 Culture:
 Begin Year: ☐ B.C. ☒ A.D.
 End Year: ☐ B.C. ☒ A.D.
 Credit Line:
 Accession #:
 Records Per Page: 25
 Sort By: Accession Number Descending

Advanced Search

Figure 6 - Layout of the Ackland’s collection search system indicating the three types of searches: simple radio, simple, and advanced.

The categories in the advanced search feature are department, classification (media type), artist/creator, title, medium, culture, begin and end year, credit line, and

accession number. While these are all distinct categories, two sets of fields are closely connected. Both the department and the culture fields primarily focus on the national and geographical origin of art. The department field focuses on broad categories like “Asia,” while the culture field lists specific nationalities and regions, like “Japanese, Osaka.” The classification and medium fields are also related in that they both are concerned with the form of a work of art. Like the department and culture fields, these fields also can be distinguished by their degree of specificity. While a user could search for “Photographs” generally with the classification field, the medium field would allow one to specify what kind of photograph the individual is seeking, such as an “albumen print from wet collodion,” instead of only its format. Of these nine total categories, two feature drop-down menus (department and classification) and four provide the option of query completion (artist, title, medium, culture) (Fig. 7). Both of these features help to reduce the possibility of users submitting a query that retrieves no results. If a user’s submission returns zero hits, this may have the effect of discouraging users from continuing to search the collection, which makes the query completion a positive feature. Especially since art collections are never truly comprehensive, even very standard searches for a particular artist may retrieve no results. Many other art museums today account for this by having hierarchical faceted search systems that directly provide users with terms to pick from rather than asking users to select their own terms since self-selected terms cannot be guaranteed to match collection holdings. While query completion and browsable facets are very useful from a practical standpoint in that they make users aware of potential search topics that will result in retrieval on a particular search system, they can complicate research of user information needs because they make it hard to decipher if a

query is truly representative of a user's needs or if it is primarily the result of one of these features.

The image displays two side-by-side panels of the Ackland Art Museum's advanced search interface. Each panel contains a series of search filters. The left panel shows the following filters: Department: (any), Classification: (any), Artist/Creator: (any), Title: (any), Medium: (any), Culture: (any), Begin Year: (any), End Year: (any), Credit Line: (any), Accession #: (any), Records Per Page: 25, and Sort By: Accession Number, Descending. The right panel shows the same filters, but with the Classification dropdown menu open, revealing a list of options: Architectural Elements, Books, Ceramics, Coins Medals, Collages, Costume, Drawings, Furnishings Equipment, Glassware, Installations, Manuscripts, Metalwork, Mosaics, Musical Instruments, Paintings, Photographs, Prints, Sculpture, and Textiles. Both panels have a Reset button and a Search button at the bottom.

Figure 7 - Detailed view of the Ackland's advanced search feature showing the department and classification drop-down menu options

Although the Ackland Art Museum's search system does not feature faceted browsing or navigation through images as is often common today for collection search systems, there are several renowned art museums that have search systems that are similar to the Ackland's in structure. The most notable comparison can be drawn with the National Gallery of Art in Washington D.C. Like the Ackland, its search interface includes extensive textual instructions. While these instructions have the potential to increase effective retrieval rates, there is also considerable evidence that users avoid opportunities to learn how to best use systems, especially by neglecting to read instructions posted online (Markey, 2007, p. 1078). The Ackland's search system also includes accession number and credit line as search categories, which are not commonly incorporated into many art museum search systems today. Other museums that feature

similar categories and drop-down menus include the University of Chicago's Smart Museum and the Detroit Institute of Arts (Fig. 2).⁶ These search characteristics diverge from those featured on other art museum collection search systems today, but this is not to suggest that these features are not widely or effectively used. The search log data will provide the evidence needed to determine the usefulness of these search categories.

Beyond the structure of the search system, the type of art objects that are in the museum's collection is also noteworthy because this will likely affect what users search for as well as what they are able to retrieve. The Ackland museum holds approximately 18,000 works of art, with strengths in Asian art, seventeenth to nineteenth-century western European painting, and European and American prints.⁷ Prints account for approximately sixty percent of the collection. While less extensive, the museum also has a strong photography collection made up of around 2,300 images. The William Meade Prince collection is the largest compilation of works by a single artist owned by the museum, consisting of nearly 1,800 pieces of art.

While the collection can only be searched by text, images are important to the collection search system. The museum received an IMLS grant in 2010 to digitize all of the works held by the Ackland; the project was recently completed in 2014 (Scott Hankins (Registrar at the Ackland Art Museum), personal communication, June 5, 2015). While not all of these images have been added to the museum's database and collection search feature yet, the vast majority of records found through the search system are accompanied by images. Several studies have found that thumbnail images are effective as access points and aid in browsing (Hastings, 1994, p. 62; Skov & Ingwersen, 2014, p. 96).

In addition to the collection search system, the museum's website also features a static collection highlights page that features some of the institution's most prized holdings (Fig 8). Featured works include Delacroix's *Cleopatra and the Peasant* (1838), Sadeler's *Triumph of Wisdom over Ignorance* (1600), and Gilbert's *Nina Cust* (1894). While not a central focus of this study, the recommendations section will note some of the most highly searched for works of art in the collection and compare them with those pictured below. Knowing what works are of particular interest to visitors, either online or in person, can help museum staff select works to put on display that have public appeal and also help inform decisions on which traveling exhibitions might be best suited for the museum's particular audience. The collection highlights page is also notable because it contains the hyperlink to the online collection search system. From the museum's home page, it takes two clicks to get the search system.

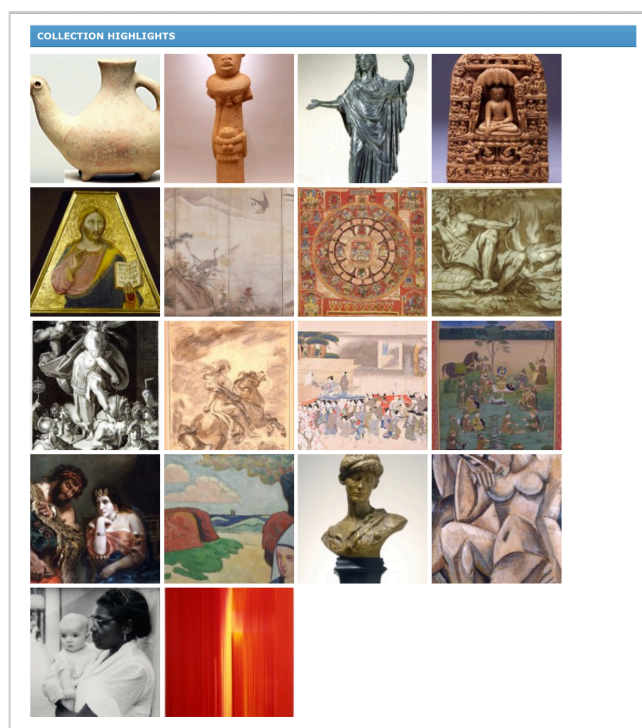


Figure 8 - Ackland's collection highlights page featuring some of the museum's most renowned works - <http://ackland.org/collections/>

This introduction to the Ackland's search system has established both what kinds of categories users are prompted to search for when using the system and the types of artwork that might typically be retrieved. While neither the search parameters nor the museum's holdings completely prevent users from searching for specific kinds of art objects or approaching search in a particular manner, they do limit the retrieval possibilities and therefore will likely have a significant influence on search behaviors. In the literature review that follows additional search categories are defined based upon human subjects research. Rather than aiming to validate search categories found on existing systems, this research attempts to classify the needs of those interested in art objects to understand their behaviors. They strive towards a user-centered ideal and follow Belkin's charge to concentrate on developing a corpus of needs.

Notes

¹ <http://uamcollection.libarts.colostate.edu/>

² <http://www.tate.org.uk/art/search>

³ <http://art.thewalters.org/browse/?type=tags>

⁴ DMA - <https://www.dma.org/collection>, V&A

http://collections.vam.ac.uk/information/information_fabricvisualiser, Rijksmuseum

<https://www.rijksmuseum.nl/en/search?f=1&p=1&ps=12>, IMA

<http://collection.imamuseum.org/results.html>, Cooper-Hewitt <https://collection.cooperhewitt.org/>

⁵ <http://britishart.yale.edu/collections/search/frames>

⁶ <http://www.dia.org/art/search-collection.aspx>

⁷ <http://ackland.org/collections/about-the-collection/> - Note that many of these values were confirmed by using the collection search system itself

LITERATURE REVIEW

Understanding the needs of the users of a particular system necessarily requires one to have some knowledge about the users. For the purposes of this study, it is essential to have a conception of who uses online collection search systems if the categories of queries they enter to retrieve information are to be assessed. It is assumed that museum employees are significant users of these search systems because they are helpful to them in performing their job responsibilities, but the search systems also have a much broader appeal. While museum collections databases “have traditionally functioned as internal documents, tailored to the needs of museums registrars and curators,” there is increasing evidence that the public actively seeks information from these systems (Cameron, 2012, p. 226).

The public that uses online art museum collection search systems cannot be definitively determined, but these individuals are likely to be associated with one or more of the following groups. Most broadly, individuals seeking visual information are often identified as users of online collection search systems. In addition, art historians, artists, and humanities scholars also have been the focus of numerous studies on both textual and visual information retrieval in the context of the Web, digital libraries, and museums. Specific analyses of art museum websites show that the majority of visitors are not art professionals, yet these individuals are still significant system users. For instance, a study by the San Francisco Museum of Modern Art (SFMOMA) determined that 18.56% of its website visitors were business and technology professionals while architects, artists, art

historians, and museum professionals each only composed between three and eight percent of users (Mitroff, 2007, para. 19). While these percentages are not indicative of museums broadly, they do show that these websites attract a range of users. In this literature review, the most recognized categorical frameworks associated with the various groups typically associated with visual information seeking will be identified and then subsequent studies that used these frameworks for analysis will be discussed. Additional research of note that does not utilize the most standard categories, as well as literature focusing specifically on the needs of artists and the contributions of log analysis to the field, will close the section.

Before the specific categories are discussed, two theoretical works will be mentioned that are essential to understanding how people understand and use images. Building off of Panofsky's three levels of meaning within art (pre-iconography, iconography, and iconology), Sara Shatford simplified these categories into two major groupings – what an object is *Of* and what an object is *About* (1986, p. 43). The author argues that all “pictures are simultaneously generic and specific” and that what a picture is *of* and *about* can be understood in both generic and specific terms (p. 47). For instance, while a painting may be *of* a lily, it may actually be *about* femininity, religion, or purity. The multiplicitous meaning of images is part of what makes it so hard for users to consistently classify them in the same way.

In addition, the research of Raya Fidel indicates the complexity of an image's essence as well as its use. Fidel developed the concept that all end uses of retrieved items are on a continuum between the Data Pole and the Objects Pole (Fidel, 1997, p. 189). Uses associated with the Data Pole involve extracting information from an object while

uses that characterize the Objects Pole entail using an image as an image. Browsing behaviors are very typical for behaviors associated with the Objects Pole because users looking for a visual image do not know which image they will pick from a set of retrieved results until they have seen the other images in the set (Fidel, 1997, p. 193). Fidel also notably remarks that art historians are unique among users because they often need to retrieve images as both objects and information sources and therefore systems designed for them cannot focus on a single type of retrieval (p. 189-190). Again, defining and understanding images in categorical terms is difficult because they can be interpreted and accessed in more than one way.

Of all the categories of image descriptions generated through user studies, Jorgensen's 12 image classes and Hollink, Schreiber, Wielinga and Worring's three image categories (conceptual, nonvisual, and perceptual) are the most prominent in the literature (Jørgensen, 1995; Hollink, Schreiber, Wielinga, & Worring, 2004). The attributes developed by Jørgensen and Hollink et al. have often been reused in subsequent research. In both the initial studies that developed these categories as well as the research that later emulated these studies, the user keywords that formed the foundation for the image classes were typically generated through description of images selected by researchers rather than through active search. While generating descriptions that are not indicative of an actual need is a potential weakness, this approach also allows users to express their needs without being limited by the constructs of an existing system. In addition, these categories have been applied to data gathered from picture databases and logs, while other studies have focused on naturalistic search to develop findings distinct from these established categories.

Of the two groups of categories, Jörgensen was the first to develop a scheme for image descriptions. The 12 category names are objects, people, color, content/story, location, description, visual elements, art historical information, people attributes, external relation, viewer response, and abstract (Jörgensen, 1996, p. 211). These can be grouped into the three larger categories of perceptual (the physical content of an image), interpretive (requires a user's intellect to perceive), and reactive (emotional response). These categories were developed in response to 82 participants' interactions with randomly selected images from the twenty-fifth edition of the Society of Illustrators Annual for Jörgensen's doctoral dissertation. Descriptions of the images by participants were the primary source of the categories, but the subjects also participated in a sorting task of the images to further establish appropriate groupings. Regardless of the task, the object class was the most frequently used by participants (Jörgensen, 1996, p. 211). The object class includes objects that can be literally perceived in an image. People, content/story, and color were also ranked highly, though their ranking differed by task.

One study that uses Jörgensen's categories is Chen's analysis of the search methods of 29 undergraduate art history majors. These students were enrolled in one of two courses on medieval art and needed to retrieve twenty images for their final term paper. The author conducted both a presearch questionnaire with the participants that collected the terms students planned to use to search for images for their paper topic as well as a postsearch questionnaire and interview that collected the actual words and phrases they used. A total of 534 queries were gathered from the two questionnaires and coded using Jörgensen's 12 image classes. The location class was the most frequently used with 22.62% of the judgments related to this attribute. Other prevalent classes

included objects (17.65%), art historical information (10.52%), and people (7.46%) (Chen, 2001, p. 270). This differs from Jörgensen's original study in which the Objects class was the most popular. In addition, the participants in Chen's study very rarely used content-based features, like color, in their queries. Finally, the queries submitted revealed that refiners, such as a specific time period or material, were rarely included in the students' search terms (Chen, 2001, 269).

Even more recently, Jörgensen's classes were used in a Master's paper by Heather Lowe that utilized log analysis to evaluate search log data released by ARTstor in 2009 (2013). While the data mostly consisted of user submitted queries and did not include the full range of fields typically associated with a search log, the total of 10.75 million queries that were submitted over a five-year period provided a wealth of information (Lowe, 2013, pp. 2, 33). Of these queries, 4,620 queries from 12 different institutions were selected to be manually coded according to Jörgensen's 12 image classes as well as the Enser-McGregor uniqueness/complexity scheme (p. 36). Of particular note were the study's findings on the use of artists' names and the changes in query composition over time. Lowe found that frequency of queries for artists' names has progressively increased over time. The study found that "Over all queries sampled, the average percent of queries judged to be artist information was 28.77% in 2005, 37.53% in 2007 and 40.26% in 2009 (Lowe, 2013, p. 59). When only art historical queries alone were considered, these percentages were even more pronounced. Queries related to artists' names went from composing 49.61% of art historical queries in 2005 and rose to 73.55% in 2009 (Lowe, 2013, p. 59). The data also indicated that the way in which queries for artists were being composed had changed. Overall, there was a shift towards using first names only to

search for artists rather than more formally using the individual's last or complete name (Lowe, 2013, p. 58). To summarize the queries as a whole according to Jörgensen's attributes, the following percentages were reported: 60.65% (art historical information), 13.83% (people), 9.913% (locations), and 8.23% (objects) (p. 64). Percentages for visual elements and color were both less than 1%.

In addition to Jörgensen's classes, Hollink, Schreiber, Wielinga, and Worring also developed a classification system for image descriptors that has been widely used. Their study aimed to determine what non-expert users look for in images (2004). This study is unique in that it used text rather than images to generate descriptors. In the study, textual selections from a children's book, an historical novel, and a newspaper, were provided to thirty participants who then used these selections in two ways. First, participants were asked to write a description of a mental image that the text generated in their minds. Second, they composed queries to find a similar image on the Alta Vista search engine. While two tasks were conducted in this study, the researchers were more concerned with the description task and claimed that image descriptors could be used as both query and indexing terms (Hollink et al., 2004, p. 605). Results from both tasks were categorized as conceptual (semantic meaning), perceptual (formal visual characteristics), or nonvisual (contextual associations) by the researchers as well as two additional reviewers to ensure consistency. Conceptual terms were used most frequently, with 87.2% of the terms fitting in this category (p. 619). Perceptual terms accounted for 11.9% of the words generated by users. These terms describe the features typically targeted by CBIR. Within this category 37.1% of the terms were associated with spatial relationships or composition and 32.1% related to color while no terms were used describing an image's texture (p. 620). Only

0.9% dealt with nonvisual descriptive information like date or creator (p. 619). These percentages differed significantly based on which textual source was used to generate the terms. Finally, the article found that when generating a query, participants used more specific terms with fewer perceptual words than when describing a mental image.

A subsequent study by Isemann and Ahmad (2011) used the categories developed by Hollink et al. and came to comparable conclusions. In this study 48 participants took a survey that asked them to provide the keywords they would use to find three representational paintings online. Twenty-one of these participants were considered experts due to their education in art history. One of these participants had a PhD in art history, but the minimum requirement to be considered an “expert” was a year of art history classes at the university level. The images described by these participants included a landscape by Monet and figural paintings by Schiele and Pallaiolo. After coding these descriptions, this study found that the conceptual level was the most frequently used with 65.6% of the terms assigned to this category (2011, p. 148). Nonvisual terms were used 25% of the time, which is a great increase from 0.9% reported in the original study by Hollink et al. This can perhaps be explained by the fact that this study included art experts who were familiar with specialized terms while the previous one was limited to nonexperts, though the difference in the stimulus for description between the two studies could have also altered this. Finally, 9.4% of the terms were perceptual. The breakdown of percentages among the perceptual subcategories was as follows: technique (50%), color (39.6%), composition (8.3%), and texture (2.1%) (2011, p. 149). Comparing the different user groups, the findings suggested that “Laypersons focus more on conceptual information, while art historians use more descriptors, that are

not directly visible in a painting and also more abstract categories” (2011, p. 150). There was not a significant difference in the usage of perceptual terms between the two user groups, though the authors question the validity of this finding.

Beyond those studies that adhere to the frameworks developed by Jørgensen and Hollink et al., others deserve mention because of their context, methods, or the unique nature of their findings. Hastings’s (1994) doctoral thesis focused on how art historians looked for paintings in a digital library collection. The collection participants interacted with was the William L. Bryant Collection of West Indies art, which contains 66 paintings. Eight art historians at the University of Florida who were experts in Haitian art volunteered for this study. The study asked the participants to interact with both an online collection of the works and physical photographs of these art objects to create queries. Because the queries generated by the participants had limited amounts of conceptual overlap, the study was primarily exploratory in nature rather than conclusive. The author concludes that “The major classes of queries in order of frequency [were] Identification, Subject, Text, Style, Artist, Category, Compare, and Color” (p. 86). As with many of the previously mentioned studies, visual characteristics, like color, were not essential to descriptions.

A study conducted by Bates, Wilde, and Siegfried at the Getty’s Art History Information Program (AHIP) is especially unique because it focused broadly on humanities scholars, rather than a particular set of academics within this category. The 22 scholars who took part in this two-year project created a total of 165 natural language statements (NLS) and 1068 queries submitted directly to DIALOG that were later analyzed upon capturing the log data (Bates, Wilde, Siegfried, 1993, p. 1). The research

study used the work of Stephen Wiberley, who did an extensive study on coding the types of terms found in seven humanities indexes, as a foundation. Wiberley (1988, p. 3) found that singular proper terms, or words that designate a particular person or creative work that can be connected to a certain moment in time and space, constituted more than half of the vocabulary found within humanities indexes. Aligning with the structure of these indexes, the Getty study found that 91% of NLSs were for a subject search, and subjects related to individuals composed 45% of the total NLSs (Bates et al., 1993, p. 15). This was the second highest subject category after common terms. A comparison between Tefko Saracevic's 1988 NSF study of science queries and the queries collected from the scholars in this study was completed. The use of proper names, such as the names of individuals and works, were completely absent from the search statements submitted by scientists in the NSF study (as reported in Bates et al., 1993, p. 16). In contrast, over half of the queries submitted by humanities scholars referred to the names of people or their creations. Because the humanities focus so strongly on the works of people, such an emphasis is logical. Less expected findings included the rare use of movements as queries ("nationalism" or "impressionism") and few occurrences of classes of creators ("painters") (Bates et al., 1993, p. 22).

Like Bates et al., Enser's study of image requests submitted to the Hulton Deutsch Collection found that known-item searching is important in image retrieval (Enser, 1993). This research determined that for users "The need to retrieve images which depict a specific person, event, location, or object is greater than that to retrieve pictures of generic items in all user categories" (Enser, 1993, p. 31). This suggests that image retrieval is focused on subjects that can be represented using proper nouns rather

common terms. In addition, approximately 69% of the queries gathered were for unique entities, indicating that image retrieval systems must be equipped to deal with a broad range of specific queries (p. 27).

While art historians, historians, humanities scholars, students, and individuals with no expertise in art have been subjects in the studies mentioned so far, one group that has been absent from the research on perception and categorization of images that one might expect to be more prominently featured are artists. Although visual artists are often conceptually grouped with art historians, in actuality they have vastly different information needs (Cobbledick, 1996, p. 347). While some may assume that artists primarily consult works from the accepted canon of art history, this assumption is not supported by research. The types of sources artists use for inspiration, visual reference, technical artistic knowledge, and marketing vary widely (Hemmig, 2009, p. 683). Rather than seeking information in art-specific institutions, like museums or art libraries, artists actually typically prefer more popular sources, like public libraries or Google (Hemmig, 2009, p. 695; Gregory, 2007, p. 63). Still, the visual perception of artists is a significant area of research. Hekkert and Van Wieringen conducted two studies to determine how art experts (industrial design and fine arts students) view works of art compared to non-experts by having both groups view works of art created by young artists in one study (1996a) and rate mechanically-manipulated works in regards to their aesthetic value in the other (1996b). Both studies found that artists were focused on analyzing a work's formal qualities while non-experts were most absorbed by representational objects (Hekkert & Van Wieringen, 1996a, p. 119). The focus on objects seems to align with

Jørgensen's findings, while the attention given to visual elements by experts could be accounted for through the use of artists as the study's subjects.

Finally, while both the Getty study and Lowe's Master's paper utilized log data, there are not a great deal of log analysis studies with which this project can be directly compared. Log analyses of art museum websites are rare and emphasize evaluation of particular systems rather than the broader needs of their users. A study conducted by the São Paulo Museum of Art Library is one of the only research projects of its kind that can be found in published literature. This library completed a log analysis in 2010 using Montalog software that evaluated a year of searches gathered from two of their database systems. For both databases, 25% of the queries retrieved no results (Di Grazia Costa, Napoleone, & Da Rocha, 2012, pp. 32, 34). Notably, it found that among the successful searches, the most common queries involved the names of artists (2012, p. 34). This suggests that search systems are properly equipped to handle these specific requests for proper names. Similar to the São Paulo Museum of Art Library's more evaluative approach, log data was studied by McLaughlin et al. using WebTrends to assess the usefulness of an online chat system called "Curator on Call" for the USC Interactive Art Museum website while also gathering some statistics on their online visitors (McLaughlin, Goldberg, Ellison, & Lucas, 1999).

More broadly, log analysis has been used for many studies on image searching in commercial web environments. While not directly applicable because of its context, Goodrum and Spink's analysis of 1,025,908 queries from the Excite search engine helps to suggest some characteristics typical of image queries. The uniqueness of these queries, when compared to textual queries, was one of the study's central findings. Rather than

suggesting categories of interest, this study indicates that queries submitted to find images typically only occur once and that they often exhibit a high degree of query modification due to their lack of precision (Goodrum & Spink, 2001, pp. 295, 303).

These various user-focused studies, set in different contexts and concentrating on a number of different user groups, provide valuable frameworks by which visual information can be categorized as well as benchmarks for the degree to which these categories have been important to users in the past. From Jörgensen to Lowe, user interest has especially focused upon Jörgensen's object class and Hollink et al.'s concept class with little attention being given to the formal elements of art. Perhaps the disinterest in visual elements is a result of the particular user groups researched in these studies. Log analysis has played a role in better understanding how searchers seek visual information, though it is typically used to analyze commercial web searches rather than those that are strictly art historical in nature. It is hoped that future research will continue to employ this research method, in addition to the more traditional interview and survey approaches, to come to a better understanding of the needs of those that use art databases and museum search systems to improve retrieval.

METHODS

Method Selection: Log Analysis and its Strengths & Weaknesses

Transaction log analysis (TLA) was selected as the primary research method for studying how users describe and search for art because of the method's ability to capture users pursuing real information needs and record the searches of exponentially more individuals than could be observed through a lab search study conducted in person (Dumais et al., 2014, p. 351). One of the shortcomings of several studies described in the preceding literature review on how users search for art images is that the data gathered was not the result of a real information need (Hollink et al., 2004; Isemann & Ahmad, 2011). Especially in studies that focus exclusively on retrieving images, there seems to be a tendency to accept an image description as the same thing as a visual information need. Unfortunately reverse engineering searches does not actually work since a user may not use the same terms to describe an image that he or she uses to describe a need for one of these items. Furthermore, it is uncertain if the image provided by researchers to be described is one the user would actually want to search for and retrieve. Log analysis circumvents these problems by providing real instances of users searching, often using their own words, for information they actually want for a specific purpose. Finally, there is no possibility of researchers influencing user behaviors because they have no actual interaction with the participants. Rather than studying the people themselves, log analysis focuses on the traces their behaviors leave behind.

Despite the strengths of log analysis as a method, it also leaves researchers with many unknowns, especially in regards to the identity of the system users and their goals for search (Dumais et al., 2014, p. 352). While differentiating how various user groups searched for art objects and their associated informational records is the focus of a considerable amount of literature (Hekkert & Van Wieringen, 1996b; Chen, 2001; Isemann & Ahmad, 2011), identifying distinct individuals on a search log is impossible without having additional information. Internet Protocol (IP) addresses can be used to identify a group of searches submitted from the same computer, but it is difficult to say for certain whether the individual querying the system is the same throughout a series of searches (Wildemuth, 2009, p. 169), even those close together in terms of time. An IP address cannot be substituted as the equivalent of a person. Therefore details about the individual submitting the search, like their expertise in the field of art history, cannot be known. Not being able to know characteristics like the age, educational background, and profession of the searcher is a notable disadvantage of the method for this study, but this drawback does not overshadow the method's other benefits.

Another weakness of log analysis is that there is no way of deciphering the purpose of a user's search and whether or not the individual's search was successful. The log data documents what terms were used to search for art records as well as when a record has been clicked on, but there is no indication provided in the log as to why certain words were used for search or whether the list of retrieved records met the user's needs. For instance, in the case of the Ackland's log it cannot be determined if a user began a search in order to retrieve an image (Object Pole) or textual information (Data Pole) and if or how this initial goal changed during search. In addition, while it is sometimes

assumed that a click on a link or object suggests genuine interest, log data does not provide researchers with the opportunity to follow up with searchers on their actions to determine if this is truly the case, except in very special circumstances in which more data than given in the log is known about the participants. Likewise, the meaning of terminating a search is unclear. While a user may decide to end a search when the desired information is found, searches may also be discontinued due to user frustration or circumstantial factors unrelated to the search itself, such as having to go to a previously scheduled appointment.

While log data leaves researchers with many uncertainties, such as user identity and search goals, it does provide a wealth of information on real search behaviors that can be gathered conveniently and rapidly. Especially with the short timespan one is given to write a Master's paper, log analysis is an effective method because it allows one to spend more time on analysis rather than collection. Logs actually provide one with so much information that it is impossible to pursue all the relationships between different variables that they present. Due to this, it is critical for researchers to identify the information within the logs that will help most with their particular research questions and for them to have specific plans for analyzing this data.

Documentation of Decisions for Data Cleaning & Analysis

In the final results user queries were the unit of analysis given the most attention, but before this analysis could begin a considerable amount of effort had to be expended to obtain the data as well as extract and organize the desired information from the data set. The staff at the Ackland Museum were extremely helpful in streamlining the data gathering process by providing the data in a usable format very quickly after the principal

investigator's initial request. Because the Ackland's online collection search system is not hosted on the museum's ackland.org domain but is instead found on the general unc.edu domain, it was necessary for staff at the Ackland to contact UNC's Information Technology Services (ITS) to retrieve the data. The search system is the only museum feature or page not found on the Ackland's website. Because ITS retains only the most current ninety days of searches, this study was constrained to using requests submitted to the system in the previous three months. The data was provided in XML format and included tags for the date and time a search was submitted, the IP address associated with the search, the search URI (uniform resource identifier) that was entered, and a summary tag (<field k='_raw'>) that included information from all of the previous tags as well as the referring page (if applicable) and the user agent string that indicates the browser used. This structure is typical of most transaction logs (Fig. 9).

```
<result offset='205'>
  <field k='_time'>
    <value><text>2015-05-04T23:38:41.406-0400</text></value>
  </field>
  <field k='clientip'>
    <value><text>68.180.228.174</text></value>
  </field>
  <field k='uri'>
    <value h='1'><text>/ackland/collection/?action=simple&amp;search=william%20blake</text></value>
  </field>
  <field k='_raw'><v xml:space='preserve' trunc='0'>68.180.228.174 www.unc.edu -
[04/May/2015:23:38:41.406 -0400] &quot;GET /ackland/collection/?action=simple&amp;search=william%20blake
HTTP/1.1&quot; 200 63901 &quot;-&quot; &quot;&quot;Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help
/us/ysearch/slurp)&quot; 0;500931;-;-</v></field>
</result>
```

Figure 9 – Example log action (simple radio query) submitted by a robot to illustrate its structure

After obtaining this data, as well as submitting a research proposal to the University of North Carolina Institutional Review Board (IRB) (Study# 15-0535), a Python program was created to parse the file, extract the data into categories, and transform its structure so that it could be more easily manipulated in Excel. This program,

along with instructions, was also shared with the Ackland Museum so that they would have the ability to analyze their search data in the future. Element Tree was used to parse through the XML structure and separate the tags into different variables.⁸ While some variables, such as IP address, required no further alterations after being parsed, others necessitated being broken down into additional variables. Special attention was given to the URI variable in order to extract the query associated with the link, determine the search type used (simple radio, simple, or advanced), and define the action type (query, results page view, or object click).⁹

After this was achieved, the final step before writing the data to a tab separated value file was to exclude all the actions that were not the result of a user actively querying the system. These excluded actions consisted primarily of robot or web crawler requests. Robots are used primarily to locate and download web pages automatically, often so these pages can be indexed (Croft, Metzler, & Strohman, 2009, p. 32). These actions were identified in one of four ways. First, the xml file was searched to find any occurrences of the word “bot,” “spider,” or “crawler” in the raw summary tag. The entire name of any robot found (e.g. Baiduspider) was inputted into the Python program for exclusion rather than simply the key search terms ‘bot’ or ‘spider’ because these strings could potentially appear in the raw field of a valid search action submitted by a user. For instance, actions that would have been incorrectly excluded if the string ‘bot’ had been used for exclusion include those with queries for “bottle shaped” and “Henry Fox Talbot.” Second, if the raw field contained the symbol “@,” indicating that an email was included in the user agent string, the associated action was also excluded. Because there were so few instances of this exception that were not already accounted for through the

first exclusion step, the complete email rather than just “@” was entered into the Python program (ex: backend@getprismatic.com).

Two additional exclusions were not made in Python initially, though the program could potentially be updated to reflect them. Google’s OpenRefine was used to identify the user agent strings that did not begin with ‘Mozilla.’ Due to the historical development of web browsers, almost all of the descriptive names of browsers today found in logs begin with ‘Mozilla’ so user agents that lack this warrant suspicion.¹⁰ Groups of actions in which this term was absent were identified by OpenRefine and then deleted (ex: MetaURI API/2.0). Finally, actions involving HEAD requests were also eliminated from the data since this kind of request would be atypical of most users who would instead generally be submitting GET requests. Both GET and HEAD are HTTP requests submitted to a web server to retrieve a web page, but HEAD requests also retrieve the time stamp of when a page was last modified (Croft et al., 2009, pp. 33, 38). Knowing this can be helpful in determining a page’s update frequency, but this additional information has no potential to help someone better fulfill an information need for art resources. While the exclusion process was very lengthy, being certain that robot or administrative actions were not included in the final data file was essential to the validity of the final research results. These exclusions eliminated 7,849 actions from the original set of 24,578 actions, leaving a total of 16,729 search actions for analysis.

Following the completion of this process, the information was analyzed using both Excel and OpenRefine. While some of the specific decisions for portions of the analysis will be not be described until the actual results section, there are some broader decisions that apply to the study as a whole that can be effectively mentioned here.

Throughout this project, focus was especially paid to the search session as a unit. For the purposes of this study, a search session was defined as the collection of search actions associated with a single IP address with less than thirty-minutes of inactivity between requests. If a user submits a request more than thirty minutes after his or her last request, this action is separated from the prior actions as the beginning of a new session. This decision to define sessions based on time is justified because even if it could be determined that the user is the same for two actions separated by this threshold, the searcher likely will have a different information need in the latter of these two actions or will approach the information need in a different manner. A session is often equated with a unique visitor even though there may be multiple sessions associated with the same IP address (Jansen, 2006, p. 419). Previous studies have set the threshold at various lengths, but several have used thirty minutes as an unofficial standard (McLaughlin et al., 1999, p. 8; Chau, Fang, & Sheng, 2005, p. 1367).

While a strictly temporal threshold was chosen for this study, session definition can be a very complex process. Standards for defining sessions are still developing and have changed greatly over time. In one of the first published log analyses of a commercial search log conducted by Jim Jansen in 1997, sessions were defined solely based on unique users rather than on time (Gayo-Avello, 2009, p. 1825). Later researchers such as Silverstein used five minutes as the threshold, while others set the maximum time limit as any searches submitted within the same day by the same person (Gayo-Avello, 2009, p. 1826). Currently work is being done to incorporate query clustering and search patterns with temporal analysis to more accurately define actions that all relate to a single information need (Gayo-Avello, 2009).

Due to an interest in how the Ackland's search system is broadly used by various users, defining sessions greatly influenced the majority of the analysis. One way of measuring interest across users was to focus analysis on unique queries in each session. While a raw count of a query term can show general interest in an idea or category, it is also possible for a single user who is particularly dedicated to a search topic or unusually active on a search system to skew the results to reflect his or her singular interests rather than the interests of searchers as a whole. Because of this, counts of query terms for analyzing both the corpus of queries as a whole and the queries associated with the advanced search method more specifically privileged unique queries per session over raw counts.

Notes

⁸ <https://docs.python.org/2/library/xml.etree.elementtree.html>

⁹ During analysis it was discovered that the progression of action types within a search session were not always logical – See the Limitations section for more information.

¹⁰ <http://webaim.org/blog/user-agent-string-history/>

RESULTS & DISCUSSION

General Characteristics of the Log Data

Using the preceding methods, the original data set was curated into a collection of information yielding many avenues for productive analysis. To begin, broad observations that apply to the entire log will be noted in order to provide the reader with a foundation to better understand the more granular analysis that will follow.

First, a timeline was created to capture the frequency and modulation of activity on the Ackland's online collection search system (Fig. 10). This timeline, created using OpenRefine, maps the total number of actions (queries, results page views, and object clicks) recorded in the log each day from February 19th to May 19th during the year 2015. There is no log data previously captured by the Ackland to determine if the type and amount of searches in the 2015 log is typical, but it is assumed that the museum website is most active during the school year because students account for the majority of their in-person visitors. As mentioned by Marty, there is some indication that online museum visitors are also physical visitors, though further research is merited to corroborate this finding (Marty, 2007, p. 339). The museum was closed for the month of June in 2015, so avoiding this uncharacteristic time period was best for assuring a more characteristic sample of user actions.

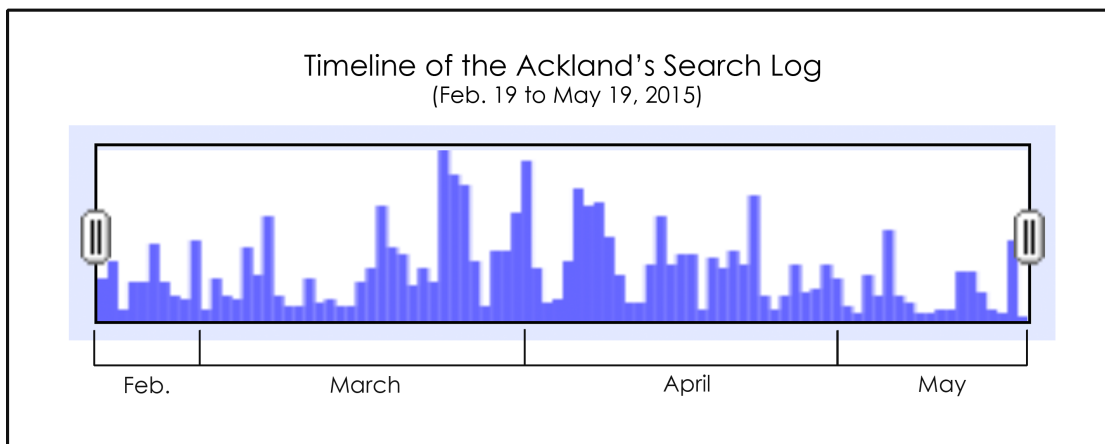


Figure 10 - Timeline of the Ackland's search log created using OpenRefine

The Ackland's timeline demonstrates, in accordance with previous studies, that the amount and type of searches can be affected by the time of the day, week, or year. For instance, for popular web search engines like Bing, Yahoo, and Google, Tuesdays typically have the most search requests, while only a quarter of the activity on this peak day occurred on Saturdays and Sundays (Taghavi, Patel, Schmidt, Wills, & Tew, 2012, p. 167). Similar generalizations can be made about the activity on the Ackland's online collection search system by analyzing the timeline. The two peak days of activity were March 24th (Tuesday) and April 1st (Wednesday). Fridays, Saturdays, and Sundays had noticeably less activity than the other days of the week. This weekly undulation aligns with other studies despite the fact that the Ackland's search system has a much more specialized purpose than those analyzed previously. The museum also has different hours of operation than is typical for most businesses as it is open to the public Tuesday through Sunday. Museum staff still generally work a typical 9-5 work week Monday through Friday. Recognizing that Saturdays and Sundays typically have lower amounts of traffic than Monday, the day the museum is closed to the public, might possibly indicate

high usage of the site by internal staff members, but no firm conclusions can be drawn from this information.

In addition to visualizing the frequency of search actions throughout the duration of the log, the data also allows generalizations to be made about the number and type of actions submitted by each “user” during this time period. Search sessions were composed of three different major actions types that were defined programmatically. These actions included submitting a query, browsing through pages of results, and clicking on art records. Using the framework discussed in the methods section, a total of 3,459 sessions were defined.

Short search sessions composed of only one or two actions were common on the Ackland’s collection search system during the time this evaluation took place. Sessions one or two actions long comprised 62.01% of the total sessions (Fig. 11). The mode for the number of actions per session for the collection of sessions was one while the median was two. There were a total of 1,250 sessions that were only one action long (Fig. 12). The brevity of these sessions has several possible implications. The session length may possibly suggest that searchers are using the system to verify information, such as the date or accession number of a painting, rather than search the collection more in-depth for purposes of discovery. The limited number of actions in each session may also indicate user dissatisfaction with the results retrieved by their queries or a complete lack of retrieved results. Some sessions may also be short because the user visited the search system due to a referring link rather than intentionally visiting the site to search the collection database itself.

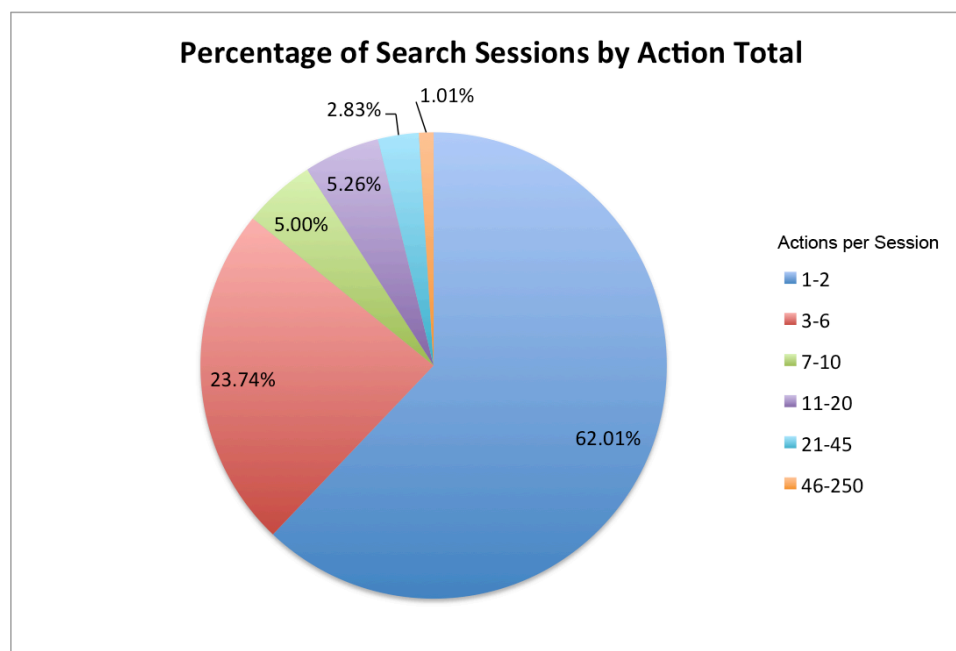


Figure 11 – Percentage of search sessions by action total showing the prevalence of sessions 1-6 actions in length

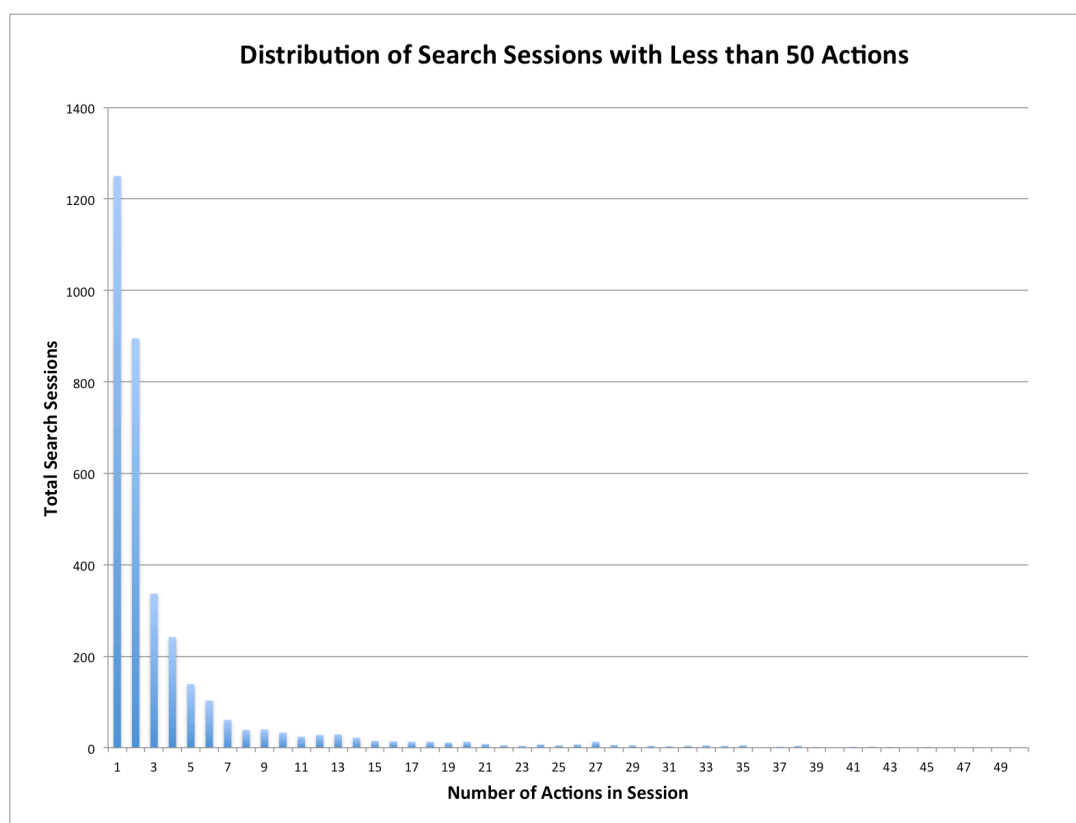


Figure 12 – Distribution of search sessions with less than 50 actions

While the majority of the sessions were brief in nature, 490 of the 3,459 sessions, or 14%, were seven or more actions long. These sessions allow further analysis in terms of how users navigate the search system and refine their queries during search. The session with the most actions lasted a total of 33 minutes and 48 seconds and consisted of 249 actions. In this session only five queries were submitted and only four of them were unique in the session. The longest session in terms of the duration of time lasted 2 hours 24 minutes and 29 seconds. This session had the second highest number of actions, with a total count of 169. A total of 11 sessions had over 100 actions. While these do not even account for a hundredth of a percent of the total search sessions, the number of actions that make up these search sessions are significant to the log data as a whole. A total of 1,572 actions occurred in this small selection of sessions, which accounts for 9.40% of the total actions in the log. While sessions are more critical to this analysis than actions because of their closer relationship to individual users, the mass of actions connected with these extensive searches warrants some attention. That the log data shows evidence of a few highly committed searchers, conforms with the results of Bates et al., who found that five users accounted for approximately 61% of submissions (1993, pp. 5, 13). As can be confirmed in Table 1, nearly half of these sessions took place outside the normal operating hours of the Ackland museum, primarily in the afternoon or evening, suggesting that members of the public could have been responsible for these submissions rather than solely staff.

used advanced							
total actions	total duration (h:mm:ss)	start time	end time	time of day	date	unique queries	search feature? (Y/N)
249	0:33:49	1:47:25	2:21:14	PM	February 28	4	Y
169	2:24:29	1:41:37	4:06:06	PM	May 18	18	Y
159	2:15:05	8:08:57	10:24:02	PM	May 6	76	Y
150	0:48:33	3:43:15	4:31:48	PM	March 26	4	Y
150	1:07:20	5:05:12	6:12:32	PM	February 20	2	Y
133	1:44:41	10:14:12	11:58:53	AM	April 23	33	Y
122	2:05:40	9:06:16	11:11:56	PM	March 7	63	N
112	0:31:23	4:10:12	4:41:35	PM	March 24	7	Y
112	1:11:51	2:27:58	3:39:49	PM	March 31	18	Y
111	0:41:00	10:54:41	11:35:41	PM	April 1	2	Y
105	1:08:45	6:25:44	7:34:29	PM	April 27	46	N

Table 1 – Duration and additional characteristics of search sessions with over 100 search actions (Note: Sessions noted as using the advanced search also often include some actions submitted using the simple search type, especially in those sessions with higher unique query counts).

The table also shows that the advanced search feature was utilized in all but two of the sessions. This demonstrates that the advanced search feature can support long search sessions. The low number of unique queries associated with some of the sessions that primarily used the advanced feature indicates that most of the actions in these sessions were results page views or object clicks, which suggests that this search type is effective for browsing. Browsing was also a common behavior in a study conducted by Skov and Ingwersen on the National Museum of Military History collection search system. Participants were especially dependent upon browsing and serendipity in finding relevant materials and were not discouraged by results with low precision (Skov and Ingwersen, 2014, pp. 96-97). This suggests that browsing is a common information-seeking behavior on collection search systems.

While the advanced search feature had high usage for the sessions with the most actions, the distribution of the three different search types (simple radio, simple, advanced) in sessions with fewer actions differs. Overall, regardless of the type of action

being performed (query, results page view, object click), the simple search method has the highest percentage of actions associated with it at a total of 38.96%, though the advanced and simple radio search types also are approaching this percentage with 31.13% and 29.91% of actions resulting from each respectively (Fig. 13). One important qualification to make about these percentages is that the simple radio search type only consists of query actions, while the other two are comprised of queries, page views, and object clicks. Following an initial query to the system using the simple radio search method, the URIs for page and record views conform to the structure of the simple search type.

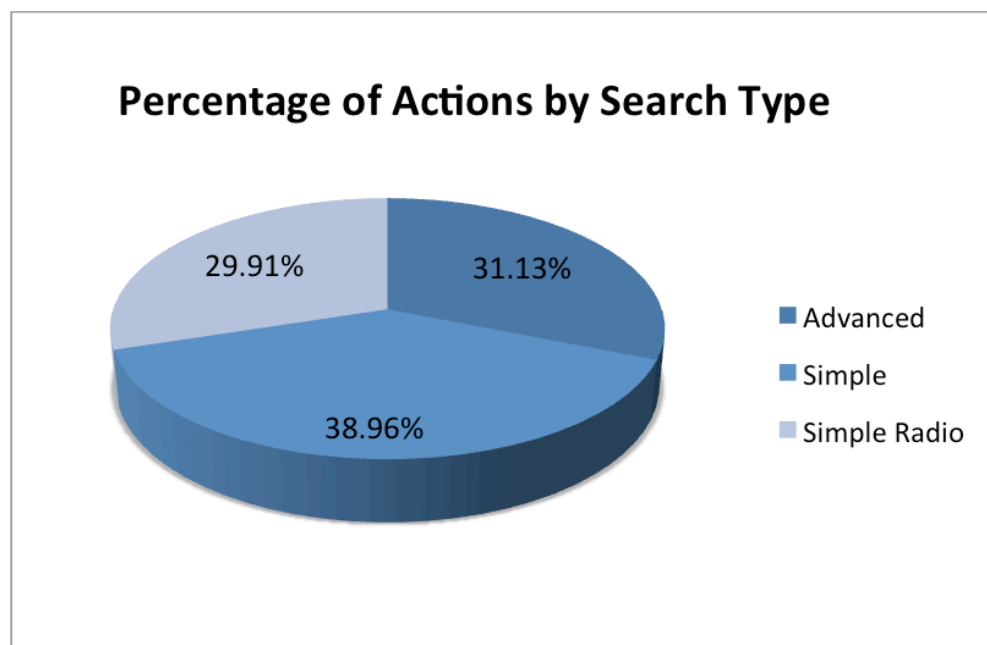


Figure 13 – Percentage of actions associated with the three search types (simple, simple radio, advanced)

Analysis of Advanced Search Category Usage

Having established some of the basic characteristics of the search sessions found within the log, these sessions and the queries they contain can be evaluated for evidence of the user information needs that initially prompted them. One of the first actions taken to begin to understand the types of informational categories that interest users most was to assess all of the actions associated with the advanced search feature. There were a total of 5,762 actions associated with this search method, which comprised just under a third of the actions found in the data (Fig. 13). This search type was singled out because the search categories selected for submission by the user were easily distinguishable in the URI. The Python program created by the primary investigator incorporated this structure into the outputted query field so that the category the search term was entered in could be easily distinguished without carefully analyzing the URIs.

While the image categories defined by Jørgensen and Hollink et al. were considered for coding, they ultimately were not best suited to this particular analysis for several reasons. To begin, using the categories that already were established by the search system had the potential to be most meaningful to the staff at the Ackland Museum and provide them with statistics that could help them improve their system in the future. In addition, having categories that could be objectively assigned made this decision both practical and scientifically sound. Since a single researcher conducted this project, there was no means of cross-checking coding schemes, like Jørgensen's, that required more subjective assignments. Finally, having gained familiarity with the types of queries present in the data through processing the xml and defining the sessions, it was felt that a coding scheme that distinguished categories within the realm of art history instead of

simply assigning terms to the discipline as a whole, as Jørgensen did for her particular data, was essential in order to define the majority of the queries more clearly.

Because the searches submitted to the advanced feature included a significant amount of duplicate queries within the same session, it was decided to only code the unique queries for each session. Limiting analysis to unique queries placed the emphasis on the session as a whole and arguably each individual searcher rather than on the queries. To be considered a duplicate, the query had to be an exact match. While many similar queries were searched for in the same session, such as “paintings AND 30 to 200,” “paintings AND 30 to 1000,” and “paintings AND 30 to 1900,” each one of these queries was considered unique. In coding, all of the categories present in a single query were counted. For the example queries above, the classification, begin year, and end year would all receive one count for each of the queries.

The results of this process found that three categories predominated in search over the others: department, classification, and artist (or creator) (Figure 14). These categories had 18.81%, 17.51%, and 16.47% of the terms found within unique queries associated with them respectively. As noted in the introductory description of the Ackland’s search system, the department and classification parameters may seem unusual to those outside of art history in their conglomeration of attributes (Fig. 7), though these categories are typically included on many art museum websites. The department field includes broad almost continental cultural groups like “Asian,” “American,” and “European” but also include the temporal/stylistic group “Modern and Contemporary” and the format-related group “Prints, photographs, and drawings.” The prominence of the department and classification categories indicate that users want to be able to use search parameters with

a broad scope that especially focus on the geographical origins of a work and their basic form. The frequent use of the artist category demonstrates that artist names are also access points of interest. That users of image retrieval systems often search by the names of known individuals is supported by the studies of Bates et al., Di Grazia Costa et al., and Wiberley. While the artist category is ranked third in terms of use when considering all of the categories, it is the first category among those that accept textual queries.

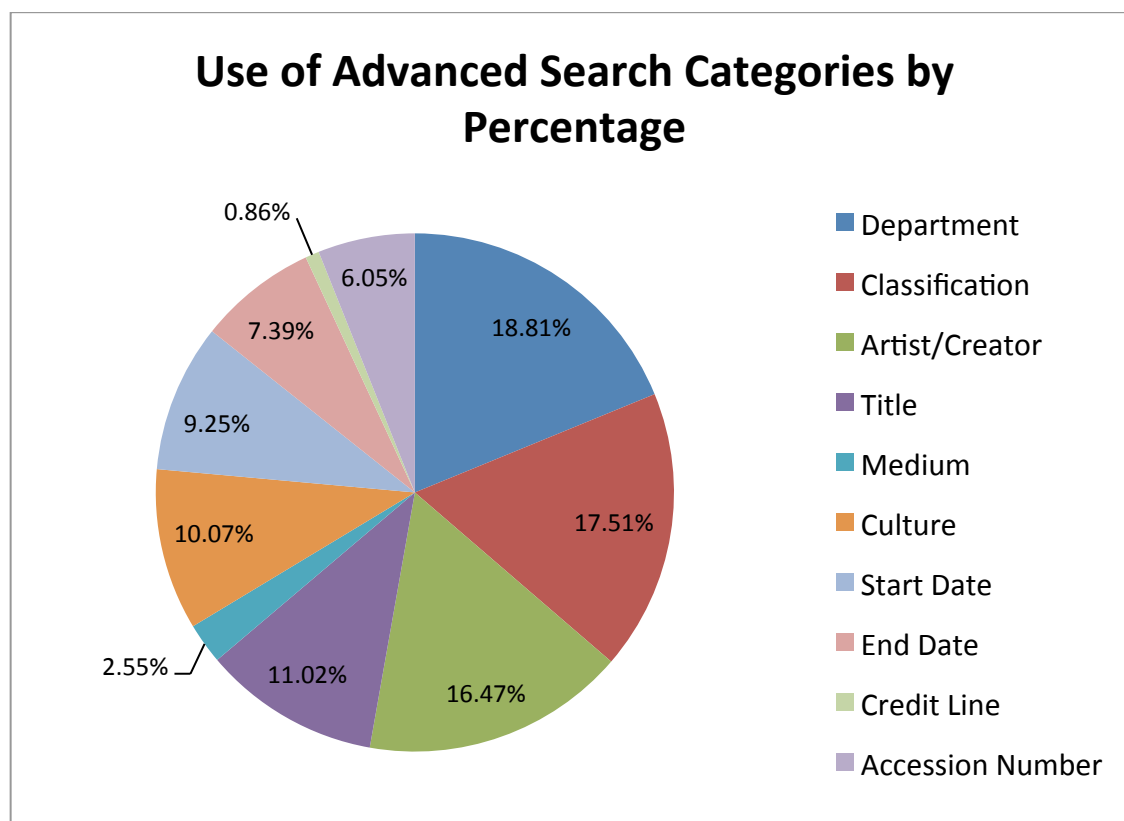


Figure 14 – Percentage of Advanced Search Categories associated with terms found within unique queries by session

In addition to identifying the categories used the most, this analysis also revealed that some categories generated little interest from users. The two categories that were least commonly used were the medium (2.55%) and credit line (0.86%). While users did frequently search using categories related to the general form of a work of art through

selecting sub-categories within the classification drop-down menu, such as “photographs” or “paintings,” the log data does not demonstrate a need to specify the medium used to construct this general form. Whether a painting was made of oil paint, watercolor, or gouache was not seen as significant to the vast majority of searchers. This suggests that many searchers may be more interested in the context, content, and form of a work of art rather than its particular materiality and the process required to create it. Other museums, like the Dallas Museum of Art (DMA) and the Columbus Museum of Art (CMA), do include specific materials as browsing keywords (e.g. “raffia” and “gelatin silver print”) and it would be informative to see if the searchers submitted to these institutions show a greater interest in material culture.

Because of the extremely specific and local nature of the credit line category, it is relatively unsurprising that it is not highly used. The credit line notes how a work came to be in a museum’s collection, whether this be through purchase, donation, bequest, or loan. It is often used to recognize individuals or corporations that donated a work of art. This category was only used in a unique query a total of sixteen times in advanced search actions. Examples of queries include “McCrindle,” “Magnum,” and “Kenan.” Since there is no query completion for this category, the user must have a considerable amount of knowledge about the collection to submit a relevant query. Museum staff may want to consider if this category is useful to have as a separate public search category. While the accession category is similar in that it does not have query completion and it is typically associated with internal use by staff, the much greater use this field had justifies its inclusion as a separate search category.

While how the design of the advanced search feature might have affected the usage of separate search categories has been briefly mentioned, more on how the structure of the search categories might have influenced the results broadly needs to be mentioned. The results suggest that there is a relationship between the order in which the categories are presented and their usage. The categories associated with the first four search boxes happen to be the most popularly used categories in search in a parallel order of descent. With the exception of the medium and credit line categories, the ranking of each category by use directly reflects the order in which the category is presented on the search system. Furthermore, the two most popular categories, department and classification, both feature drop-down menus for query submission rather than textual entry. Because this kind of submission requires less cognitive load, it is not surprising that they are highly used categories. The four search options following these drop-downs do feature query completion, but a drop-down menu expects even less knowledge of a user than query completion. It would be informative to see if the ranking of the categories changed if the order and mode of search was altered.

In order to better understand the relationship between the complexity of the queries and the system's structure, a brief analysis of the maximum number of categories employed in each search session was also conducted. It was found that users only selected one advanced search category more than half of the time (Fig. 15). Still, multiple categories were selected by a fairly high percentage of users, with 27.5% of sessions including a search that employed a search with two of the categories. The department and classification drop-down menus were frequently used together for these queries that involved two categories. Despite its seemingly repetitive nature, a very common query in

the log that is an example of this is “Prints, Drawings, and Photographs AND photographs.” While no users came close to using all ten available search categories, several users did utilize up to five categories during a search session (1.3%) .

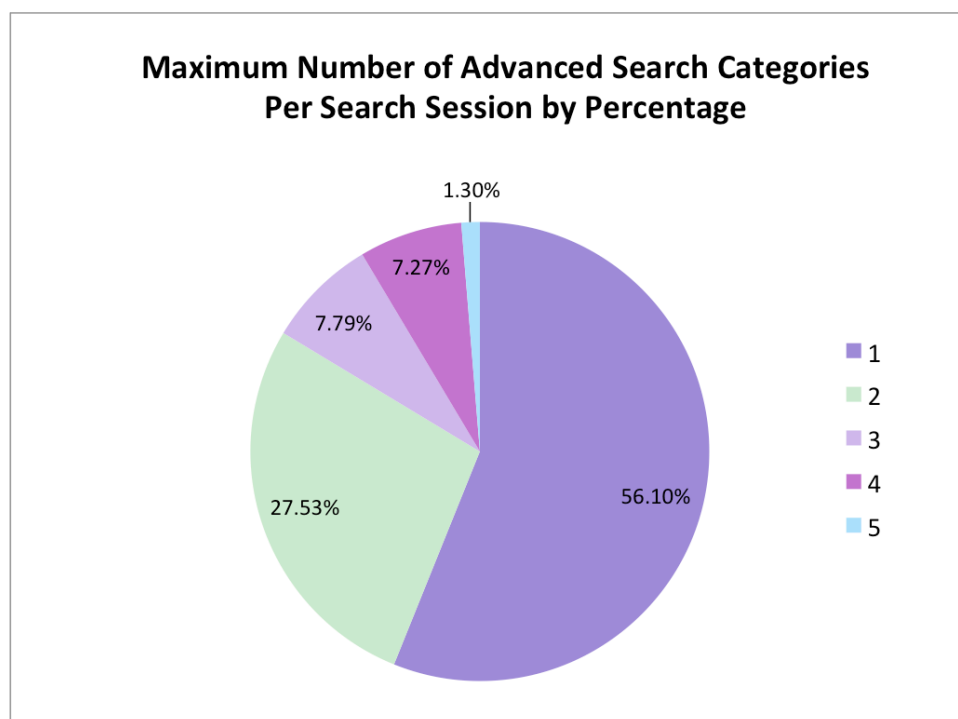


Figure 15 – Maximum number of advanced search categories used per search session by percentage

Corpus of Queries

In addition to the broad categorical analysis of advanced queries undertaken, attention was also given to particular queries themselves and the frequency of their use by visitors to the system. Because the advanced search categories are strongly focused upon eliciting solely art historical queries, considering queries submitted to the other two search types (simple and simple radio) was essential to constructing a more complete picture of user needs. To begin this analysis, it was first important to establish a corpus of queries. OpenRefine’s text facet feature was used to rapidly identify the unique query terms in the data and their frequency. In order to ensure that structurally similar terms and

their frequencies were not considered separately, case folding and stemming was used to make sure that the prevalence of certain terms was not understated. Stop words, or high frequency strings like “a,” “and,” “the”, and prepositions, were also disregarded when considering if queries were similar enough to combine. The consolidation of some of these similar terms could be achieved in OpenRefine, but a substantial amount of exceptions remained to be completed by hand following this automated process.

During this process, several guidelines were created and observed. In stemming, plurals and gerunds were always folded together into one query (e.g. “sleeping” was combined with the query “sleep”). In addition, words with misspellings with an edit distance of two or less were also combined (e.g. for the term “picasso” all of the following misspellings were accepted – “piasso,” “picassso,” “piccaso,” and “Piscasso”). Accents and other punctuation present in the query terms were also disregarded (e.g. “francois boucher” and “FRANÇOIS BOUCHER” were considered to be the same search term). No attempt was made to combine cultural and geographical terms with similar roots (ex: “china” and “chinese” were left as separate terms). The presence of ASCII encoding in the original query terms extracted from the URIs complicated the process of term consolidation because whereas some encodings were easily recognized and replaced (e.g. %20 for a space) other encodings were more rare (e.g. %F1 for ñ) and the appropriate characters for these less typical encodings were not substituted before beginning the folding process. Adding a function to replace this encoding with appropriate characters in the initial program would be beneficial.

This process highlights some of the limitations of standard text processing. While stemming and case folding do successfully bring together terms that are typically both

structurally and semantically related, these methods are incapable of identifying terms that disparate in form yet related in meaning. For instance, even search terms that can be identified by some as referring to the same concept or item, such as “durer” and “albrecht durer,” are defined as separate strings with standard text processing and the frequency occurrence of each is counted individually. Even though the terms are related, there is no indication of this in the final frequency counts. While these two search queries can potentially be identified as having some connection, in many instances the relationship between terms can be hidden.

For example, in the Ackland’s log there are a number of queries that at first seem to be isolated from the rest of the search terms in meaning, but, with enough knowledge of either the collection or other queries, can become less detached. The search term “harlot” is one of the top 50 search terms in the log. At first it seems like this search is limited to looking for works that literally represent “fallen” women, but it becomes clear that at least some of these searches are looking for a specific work due to title searches in the log for works that contain the word “harlot.” In addition, some users actually connect the keyword “harlot” with a particular artist. There are ten unique session occurrences of the search phrase “hogarth harlot” in the log. These queries indicate interest in a series of engravings titled *A Harlot’s Progress* by William Hogarth. The problem is that while this relationship may have been revealed through the information in the log, many others will remain hidden. Also, a person can search for the term “harlot” without having any knowledge of or interest in this specific set of works within the collection. Therefore including related terms like these in count frequencies is unfairly privileging them. A

more unbiased approach is to only use terms that can objectively be recognized as being related through their structure.

After text processing on the set of queries was completed, a total of 3,427 queries were identified. Both the total number of actions (Table 2) and the total number of unique queries per session for each query were analyzed (Table 3). Some queries, like “paintings” were associated with a high number of total actions (335), but a much fewer amount of unique queries (18). Both counts are significant, but this study especially aimed to determine the general interests of all users rather than the specific interests of particular users, so the unique query count by session was privileged. Comparing the types of queries found in the two tables reveals that certain search categories consistently have many actions associated with them but are found in a very small number of sessions. This is particularly true of time-related queries (e.g. “1900 to 1950”), which are completely absent from Table 3 but account for 11 queries in Table 2. This demonstrates that queries that include years are often used for browsing and that users are rarely interested in searching for works associated with the same exact timespan.

Processing these queries also revealed some general tendencies in how queries were often structured and formatted. Unlike Lowe, who found that searchers often looked for artists by their first name, the queries analyzed in this study demonstrate a preference for using only the last name of the artist (Lowe, 2013, p. 58). In the list of queries that appear in the most sessions, queries that include only the artist’s last name consistently precede queries for an artist by their full name. For instance, while there are 68 sessions that use the query “picasso,” there are only six sessions that include a search for “Pablo Picasso.” Considering the top 200 queries, “charles,” is the only search relating to an

Queries Associated with the Most Actions					
Rank	Action Count	Query	Rank	Action Count	Query
1	335	painting	35	40	falls of tivoli
2	251	damocles	36	39	kunisada
3	244	Chinese Art	36	39	virgin and child
4	229	European AND paintings	37	37	bohrod
5	172	japanese	37	37	lion
6	158	sculpture	37	37	Neck Amphora
7	145	1900 to 1950	38	35	degas
8	142	Asian	38	35	duchamp
9	126	woodblock	38	35	gold
10	125	portrait	38	35	pissarro
11	120	Modern and Contemporary	39	34	European AND sculpture
12	110	Prints, Drawings and Photographs AND photographs	39	34	Greek
13	108	photographs	39	34	William Meade Prince
14	103	mughal	40	33	apollo
15	88	paintings AND 1900 to 2015	40	33	Asian AND Chinese
16	83	picasso	40	33	seated girl holding a blossom
17	82	Ancient Mediterranean and Middle Eastern	40	33	Wharton Esherick
18	81	French AND 1700 to 1900	41	32	Ando Hiroshige
19	78	European	41	32	British AND 1700 to 1900
20	77	women	41	32	interior of the oude kerk
21	66	charles	41	32	madonna
21	66	Prints, Drawings and Photographs AND 1550 to 1650	41	32	Modern and Contemporary AND ceramics
22	64	Indigenous Americans	41	32	Oil on canvas
22	64	Modern and Contemporary AND paintings	42	31	2014
23	63	vishnu	42	31	Asian AND Japanese
24	62	african	42	31	italy
24	62	italian	43	30	European AND paintings AND 1600 to 1899
25	58	durer	43	30	goltzius
25	58	porcelain	43	30	indian
26	55	juno	43	30	prayer mat
27	54	American	43	30	Sodom
27	54	Carrick	44	29	1500 AND 1600
28	49	blue	44	29	Cuban
29	48	William Meade Prince	44	29	glass
30	47	earthenware	44	29	Italian AND 1300 to 1650
31	46	buddha	44	29	warhol
31	46	European AND Italian	45	28	madame
31	46	franz marc	46	27	family
31	46	india	46	27	white
32	44	Landscape	47	26	Asian AND paintings
32	44	Lekythos	47	26	battle
33	42	dance	47	26	paintings AND 1950
33	42	dance in a garden	47	26	sellaio
33	42	Prints, Drawings and Photographs AND 1950	48	25	2010
33	42	sadeler	48	25	gallery 15
34	41	Chinese	48	25	Harlot
34	41	hiroshige	48	25	Ramayana
34	41	japanese print			

Table 2 – Queries associated with the most actions (query, page view, object click)

Queries submitted in the most sessions					
Rank	Query Count	Query	Rank	Query Count	Query
1	160	damocles	20	12	India
2	65	mughal	20	12	Modern and Contemporary
3	61	Prints, Drawings and Photographs AND photographs	20	12	pissarro
4	57	sculpture	20	12	rose piper
5	50	portrait	21	11	blue
6	45	woodblock	21	11	japanese
6	45	picasso	21	11	Lot and His Family Fleeing from Sodom
7	44	charles	21	11	narcissus
8	31	Neck Amphora	21	11	rubens
9	30	juno	21	11	weenix
10	29	Lekythos	22	10	American
11	25	duchamp	22	10	hogarth harlot
12	22	Carrick	22	10	osei bonso
12	22	European AND paintings	22	10	standing buddha
13	19	buddha	22	10	watch
13	19	falls of tivoli	22	10	William Meade Prince AND 62.27
13	19	franz marc	23	9	African and Oceanic
14	18	Asian	23	9	blue mountain and lake
14	18	apollo	23	9	bull
14	18	painting	23	9	cleopatra and the peasant
15	17	dance in a garden	23	9	European
15	17	gerlovin	23	9	Henri Rousseau
15	17	Indigenous Americans	23	9	hiroshige
15	17	italian	23	9	seated girl holding a blossom
15	17	sadeler	23	9	stigmatization of st. francis
15	17	tobacco	23	9	sword of damocles
16	16	dance	23	9	tiger
16	16	gallery 15	23	9	virgin and child
16	16	italy	23	9	women
16	16	madame	24	8	American AND paintings
16	16	vishnu	24	8	Cuban
17	15	bronzino	24	8	Henry C. Pearson
18	14	bohrod	24	8	Jean Baptiste Oudry
18	14	degas	24	8	prayer mat
18	14	interior of the oude kerk	25	7	african
18	14	madonna	25	7	Ancient Mediterranean and Middle Eastern AND sculpture
18	14	sellaio	25	7	cat
19	13	aaron bohrod	25	7	durer
19	13	cleopatra	25	7	egyptian
19	13	hans thoma	25	7	European and Italian
19	13	kylix	25	7	Greek
19	13	photograph	25	7	harlot's progress
20	12	Ancient Mediterranean and Middle Eastern	25	7	paul valadez
20	12	Harlot	25	7	rousseau

Table 3 – List of the top queries by frequency that are unique in a search session¹¹

individual that does not include a surname. The query completion feature for the advanced search offers suggestions for first name only in the artist field in addition to options in the form of “lastname” or “first name last name”, so it should not be biasing how users are formulating queries for artists. In addition, users frequently included diacritics as well as foreign language terms. Some search sessions showed that users transitioned from entering queries in French or Spanish to English (e.g. *stele de princi* → *stele of prince*).

Key word searches for representational subjects found in a work, rather than exact title searches, were also prevalent. The work *The Sword of Damocles* (1812) had 9 search sessions that mentioned it by title, but the keyword “damocles” was associated with 160 search sessions. This search term was by far the most popular in the log, having been used in nearly 100 more sessions than the second most popular term “mughal.” In addition to the previous tables, the popularity of “damocles” as a search term is also represented visually in a word cloud (Fig. 16). The inclusion of the queries “Harlot,” “Cleopatra,” and “dance” in the word cloud also supports the claim that users often search for representational subjects within a work. The term “Cleopatra” may be specifically connected with the work *Cleopatra and the Peasant* (1838) found within the Ackland’s collection, but it could have also been submitted by the user in order to find any representations of this person instead of a singular representation. Regardless of whether these example terms can be tied to specific works, identifying that users have a tendency to search by representational subject indicates that this is an important category that should be included in an art record. For works with minimal metadata, the title of a piece may become especially important because it could be one of the few fields with

searchable text that is included in the record. When pieces of art do not have given titles, naming them based on the subjects they represent could improve retrieval.



Figure 16 - Word cloud of the top 49 search terms in the search log created using Tagul. Each string is associated with at least 12 unique queries per session in the log. The case and number of each string represents the way the query was most frequently submitted by users rather than the tokenized version of the query. Queries that include “AND” were submitted exclusively using the advanced search feature.

An additional finding of interest related to the top queries was the extent to which external links were used to find works on the Ackland’s search system. The query found in the most URIs was a keyword search for “damocles.” While structurally the link in the log for these actions was a query, many of the “searches” using this keyword involved users clicking on a link on an external website rather than actually typing in the query. In the case of *The Sword of Damocles* (1812) (Fig. 17), eighty-five percent of the search

sessions that included the query “damocles” began as referrals from Wikipedia. This helps to explain how this query came to be found in so many more search sessions than any other query found in the log. Another one of the top ranked “queries” that also resulted from external links from Wikipedia was “William Meade Prince AND 62.27.” As the only query that included an accession number in the top 200 queries, its popularity was at first confusing. It seemed unlikely that many individuals outside of museum staff would know the accession number to compose this query. That the query’s high frequency was partially the result of referrals rather than searches helped to explain this peculiarity. These two examples show the depth of information log analysis can provide concerning a search and also indicate the importance of utilizing this information to better understand user actions.



Figure 17 - Richard Westall. *The Sword of Damocles*, oil on canvas, 1812, Ackland Art Museum. The most popular keyword search in the log, “damocles,” is related to the title and subject matter of this work.

In addition to extremely popular queries like “damocles,” the table of terms that appear in the most search sessions (Table 3) also provides insight into the demand for the incorporation of new search categories, like color and abstract concepts such as emotions. A query for a color, “blue,” is ranked as one of the terms with the top session frequencies in Table 3 with a total of eleven session occurrences. Other color-related terms and their session frequencies in the log include “red” (3), “black” (2), “green” (2), and “white” (1). Summing these values, a total of 19 sessions included searches solely for a color’s name. No queries were submitted for specific colors like “turquoise,” which fall outside the standard grouping of primary and secondary colors. The names of colors were also often used to modify terms (e.g. “black cat”), but these submissions are not considered here. While color queries could show that users are looking for works that include these colors, it is important to note that these searches alone cannot definitively confirm whether users are interested in color as a formal element or if they are searching using these terms for other reasons. For instance, many works include color words in their titles (e.g. “blue mountain and lake”) so it is possible that these searches are an indirect way of completing a title search for a specific work rather than searches for the appearance of certain colors broadly. With queries like “black” and “white,” it is also potentially possible that users are searching for racial categories. The Ackland’s log data shows that there is some interest in color, but it is limited.

Finally, a number of terms that dealt with the museum space or staff were surprising to uncover. While the queries for staff names, such as “Peter Nisbet” and “allmendinger” can be seen as the result of misusing the simple radio search method by remaining on the collection button rather than clicking on the Ackland.org button, others

could have potentially been intended for the collection. Most notably, there were a total of 53 searches that involved a query for a specific gallery room. The most popular query that matches this format had a total of 16 unique session queries and was for “gallery 15.” Other galleries searched for include “southeastern asian gallery,” “gallery 12,” and “study gallery.” This suggests that users are interested in getting a sense for how art objects are organized within the museum space, or at least with the existence of different rooms within the institution. While it seems likely that staff submitted at least some of these queries and the number of queries is not comparable to those found in other standard categories, like artist, they should not be discounted. As mentioned in the introduction, museums like the LACMA and the Tate allow online visitors to the collection search system to look for objects based on their location within the museum. Although museum location may not usually be considered a search category integral to the essence of an art object or its representation, if users are looking to access works in this way they should have a means of doing so.

Query Reformulation

Through evaluating the queries submitted to the Ackland’s search system, some of the more commonly used categories have been determined. While these categories alone are significant, they can also be used to further develop knowledge about how users search for information. Beyond uncovering the queries themselves, one of the unique advantages of a log analysis is the opportunity it presents to evaluate the transitions among the queries that occur in a session. How queries change over the course of a search session is referred to as query reformulation. While analysis often centers on the query itself as a unique unit, in this study attention was especially paid on how the

broader category associated with a query changes between actions. Rather than identifying characteristics of change unique to words at the query level, such as misspellings or synonym matches, higher categories associated with the queries were privileged. Because online art museum collection search systems often suggest related items for users to explore after the initial submission of a query, it was felt that a categorical focus would be of the greatest use. For instance, if a user submits a search using a particular artist's name and clicks on one of the creator's works, what is the user most likely to want to click on next? Currently museums that have a related records feature usually suggest works by the same artist as the work the user clicked on, but there is not any evidence that confirms that these suggestions are what the user is mostly likely to want to see next.

To examine this, 57 sessions were systematically selected from the data by choosing every tenth session with six or more actions. Each of the queries found within these sessions was classified into one of the following categories: artist, culture, work, representational subject, non-representational subject, artistic form, and time. These categories were arranged hierarchically in order to aid with defining transitions between them in the log actions. The transition categories were defined as generalization, specification, parallel movement, and nondefined. The most distinct category in the hierarchy was a work of art. A query that was categorized as a work of art either exactly matched the title of a work of art or a work's accession number. A movement from any of the other categories to a work of art was considered a specification. Likewise, a transition from a work of art to one of the other categories was always coded as a generalization. The artist category is one level in the hierarchy higher than a work of art, because a single

artist may have made multiple works, and it is the only category in the secondary level of the hierarchy that had a broader category associated with it. Culture is considered a broader category of artist since multiple artists can all come from the same cultural background. Movements that were considered nondefined mostly included transitions from a culture to a form (e.g. “Ancient Mediterranean and Middle Eastern” to “glassware”). Because the range of art works associated with either of these categories is broad and variable, they were not able to be defined hierarchically in relationship to one another.

The results of coding the changes that took place from one query to another revealed that parallel movements made up the highest percentage of transitions (Fig. 18). Referencing the Web and the prevalence of hypertext navigation, Peacock et al. note, “Moving sideways can be equally as effective as scaling a hierarchy of information” (2004, p. 15). That there were 42% more horizontal movements than vertical transitions in the sample of sessions analyzed supports this statement. The most common parallel movements were artist-to-artist transitions, with 50.17% of these movements involving this type of reformulation. These moves primarily consisted of movements between different artists, rather than multiple queries submitted in succession for the same individual. A total of 76.67% of the artist-to-artist transitions involved movements between different artists. The majority of query pairs that were submitted to search for the same artist primarily involved changing the form or spelling of the artist’s name from the first query to the second query.

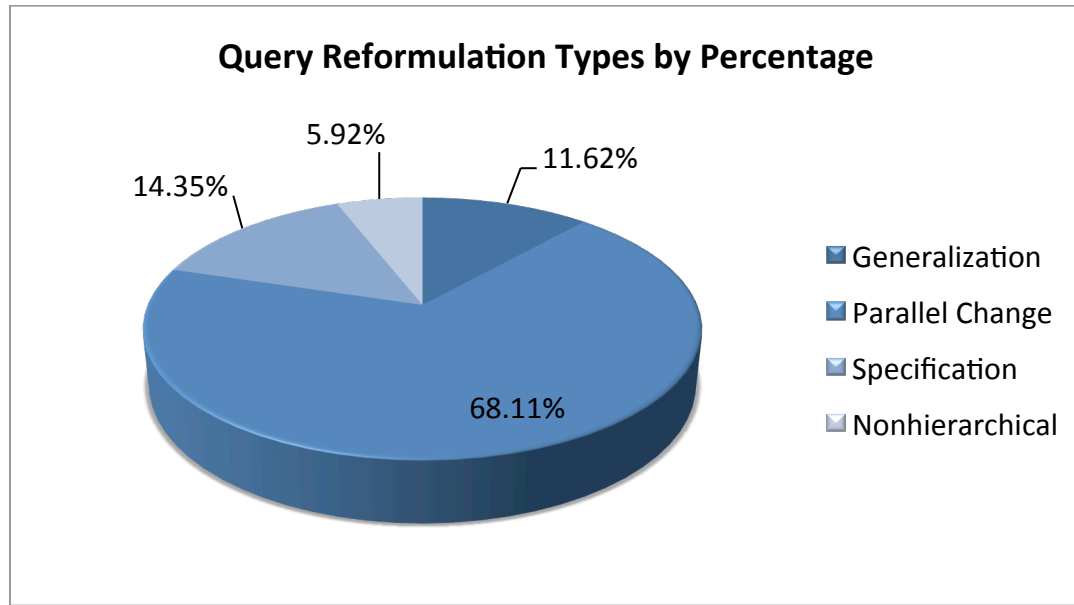


Figure 18 – Categorical query reformulation types by percentage

In order to better understand the broad reformulation types, one example from the set of 57 selected sessions will be included here. In the diagram below, the words separated by arrows are the queries submitted by users in the form they were entered. The letters and numbers in the brackets were not part of the user's original queries and are only included for clarification. The "P" stands for parallel change, which is the type of reformulation taking place between all of the query pairs in this example. The number counts the total transitions and notes each transition's place within the session.¹²

Pissarro → [P1] Gustave Courbet → [P2] jacopo amigoni → [P3] nicholas lancret
 → [P4] lancret → [P5] de witte → [P6] jacob duck → [P7] de Ruysdael → [P8]
 rubens → [P9] bronzino → [P10] pissarro

This session was chosen because it shows the prominence of parallel movements among users of the Ackland's search system as well as how frequently artists' names were used as queries. It demonstrates that users do not commonly repeat the same names

as queries, but instead submit queries for unique individuals. While the ten parallel changes between artists' names may seem extensive, another one of the selected sessions included forty parallel transitions involving artists. The actions documented in these sessions may indicate attempts by users to see if the museum has any works by famous artists, or simply artists that interest them. Perhaps because artists are typically associated with many works, they act as access points that are somewhat likely to retrieve results from the system. While using culture or the name of an artistic school as a category would generally return more results than an artist's name, it seems that users sometimes prefer more specificity than this search category offers. Rather than starting a search with the categories that have the broadest scope and then narrowing down the submitted queries after reviewing the results, users are more comfortable composing queries at their desired level of specificity and simply entering many queries of the same type to get the information they want.

Notes

¹¹ In addition to these textual queries, there were a total of ten sessions that included a null query, or a query without any entered text or selected fields. While these simply could be errors on either the part of the system or the user, they also potentially suggest that either users are unsure of what an appropriate query is to submit to this system or that they want to browse rather than search the collection. That museum visitors may have uncertain information needs is supported by Skov and Ingwersen's survey at the National Museum of Military History which indicated that 29.5% of the web visitors that came to the collections database page were not looking for anything in particular (2014, p. 94).

¹² This diagram was closely modeled off of the query reformulation diagrams found in Rieh and Xie's 2006 article "Analysis of multiple query reformulations on the web" included in the bibliography.

CONCLUSION

Through the use of search log data, this study has aimed to provide greater understanding of the types of search categories that are important to individuals who use the Ackland Art Museum's online collection search system. A Python program was written to process the log data and Excel and OpenRefine were used for analysis. Through processing, the 16,729 actions that composed the initial data file were grouped into 3,459 search sessions using a 30-minute time threshold between actions as the primary session delimitator. While 62.01% of the search sessions were two or fewer actions in length, eleven sessions exceeded 100 actions. Both quantitative and qualitative approaches to analyzing the data were used. When possible, the percentage that a particular search category was used was calculated. In other instances it was just as important to highlight some of the idiosyncrasies within the data that could not be described quantitatively. The importance of Wikipedia referrals to the widespread use of the search term "damocles" is one example of this. While these referrals cannot be classified as user-submitted queries, they still represent a significant access point to the collection and therefore are worthy of consideration along with the common search categories.

Throughout this study, user queries and the sessions they were located within were the central units of analysis. Queries unique to a particular search session were especially privileged in order to acknowledge the significance of the submissions of each individual user. Three separate, though strongly interrelated, analyses were completed to

come to a better understanding of the types of search categories that best met the needs of users. Through studying the queries associated with the advanced search feature, it was revealed that the department, classification, and artist fields are the categories most commonly used, with each accounting for 18.81%, 17.51%, and 16.47% respectively. The categories used the least were credit line (0.86%) and medium (2.55%). Overall, the advanced queries demonstrated that users do have considerable interest in broad categories associated with culture and form as well as in the creators of works of art, but not in an artwork's materiality.

The frequency of unique query terms by session across all search types also showed an interest in broad artistic forms, like prints and photographs, while further emphasizing the use of artists' names in search and introducing representational subjects as an additional significant point of access. In addition to "damocles," the prevalence of representational subjects in this analysis can be seen through the wide usage of queries like "buddha," "juno," and "apollo" in the log (Table 3). The popularity of these kinds of queries provides further support for the high percentage of queries associated with Jørgensen's object and people classes found in earlier studies (Jørgensen, 1996; Chen, 2001). That the queries are often aimed at retrieving images of specific individuals also supports the findings of Enser (1993, p. 31).

Finally, through studying the categories of queries across a session, it was found that individuals are most likely to use one category throughout the course of a search. The selected set of sessions studied showed that 68.11% of the transitions between queries were parallel changes that involved two queries from the same category or queries that were at the same hierarchical level. Of these transitions, 50.17% involved moves from

one artist to another. While the names of artists are not the most prominent of the search categories, they were consistently identified as a significant access point across all three analyses conducted in this study. This aligns with research conducted by Wiberley (1988), Bates et al. (1993) and Lowe (2013).

These specific findings are not generalizable to the search behaviors of users of online art museum collection search systems broadly, but it is hoped that they will be of use to the Ackland Art Museum in increasing access for their particular set of users. Furthermore, while this study is limited in its scope, it does provide needed information on how real users search for art records. If enough local studies are conducted on specific search systems to answer some of the research questions that remain, a broader picture of the categories and methods used to retrieve art historical works can be formed.

Recommendations

All institutions with search systems have the opportunity to analyze how their systems are being used through gathering and analyzing data that is accumulated through its daily use. Collecting this information on a regular schedule is a good practice to exercise. Especially in the case of the Ackland, in which log data is only stored for the preceding ninety days, planning is necessary if this information is to be retained. While the museum may not have the staff or resources to allocate to manage the collection and care of this data or its analysis currently, a conversation should be had on whether staff feel this information is valuable to the mission of the museum.

If the Ackland's search system is either updated or replaced in the coming years, restructuring the way in which the URIs are created so that the links are persistent would be helpful for the museum and its users. Currently there are several links on the web from

Pinterest, LearnNC, and Wikipedia to records on the Ackland's search system that are broken. While these links could positively promote interest in the museum's collection if they functioned properly, many of the links currently produce the message "Your search returned 0 results" or "The object could not be found" instead of generating a list of art records.¹³ Although all of the art records in the collection have an option to create a persistent link using the accession number, there are still many users that do not take advantage of this feature. In total, there were 116 instances in which one of these persistent links was the referring page for a search action.

If links to particular objects or lists of search results cannot be made persistent, it would be beneficial to make them at least identifiable with a particular record upon analysis. Actions associated with clicking on an art record were among the most problematic to interpret because they were not persistent and the object ID found in the URI was a random number rather than an intelligent identifier. While a URI could be categorized as a record click based on its structure, there was no way to identify which particular record was clicked on in that action. Being able to identify the works that have been clicked by users would greatly increase the value of future analysis. Rather than focusing on interest in known items by looking at search queries, being able to identify which records were clicked on the most would have provided the Ackland with information on the records that generate the most interest in the search process regardless of whether finding these records was a goal of the user's initial search.

While the structure of the system made it impossible to trace which specific art records had been clicked on and the number of clicks they received, some indication of the interest users have in specific art objects can be generated through analyzing the

queries present in the log. The sizing of the artworks in the image below (Fig. 19) takes into account unique searches that exactly match a work's title or accession number as well as keyword searches for non-adjectival words that appear in the work's official title. Keywords, while not necessarily directly connected to a particular work, were included because there was little difference among works based on exact title or accession number matches. Because *The Sword of Damocles* (1812) has already been featured at length and generated so much more interest than any other piece, it was excluded from this visualization. *Cleopatra and the Peasant* (1838) is the only work currently featured on the Ackland's static collection highlights page (Fig. 7) that is also found in the image on the next page.

In addition, work should be done to better understand the sequence of actions and the way searches are logged in order to improve future studies. The progression of actions in the log data was not always logical, suggesting that either the search system was generating URIs in an unconventional manner or that some actions were not being logged correctly. This issue seemed to particularly affect the first action of a search, which generally consisted of the query. The query from the initial action could be found in the subsequent actions, but it forced the primary investigator to do more manual counting than would have otherwise been necessary. Some of these unlogged initial actions could possibly be accounted for as outside referrals, but none of them had a referral link logged to support this conjecture. Information and Technical Services (ITS) on campus as well as the museum were contacted about this issue, but it was unable to be resolved.



Figure 19 – Works in the Ackland’s collection with the highest number of search terms related to them(excluding *The Sword of Damocles*). The titles of the works in order of their “popularity” in the log are 1. *Vessel (Neck Amphora) with Apollo, Leto, and Artemis* 2. *Juno Beseeching Aeolus* 3. *Virgin and Child with the infant of St. John the Baptist* (1560s) 4. *Virgin and Child* (1415) 5. *Cleopatra and the Peasant* (1838) 6. *Falls of Tivoli* (1807) 7. *Egyptian Princess with a Musical Instrument* (1360-1350 BCE) 8. *Dance in a Garden* (1730s) 9. *St. Jerome in Penitence* (1515) 10. *Spanish Dance* (1885) 11. *Seated Girl Holding a Blossom* (late 17th c.) 12. *Mass of St. Gregory* (1550) 13. *The Banks of the Oise, Near Pontoise* (1876).

In addition, work should be done to better understand the sequence of actions and the way searches are logged in order to improve future studies. The progression of actions in the log data was not always logical, suggesting that either the search system was generating URIs in an unconventional manner or that some actions were not being logged correctly. This issue seemed to particularly affect the first action of a search, which generally consisted of the query. The query from the initial action could be found in the subsequent actions, but it forced the primary investigator to do more manual counting than would have otherwise been necessary. Some of these unlogged initial

actions could possibly be accounted for as outside referrals, but none of them had a referral link logged to support this conjecture. Information and Technical Services (ITS) on campus as well as the museum were contacted about this issue, but it was unable to be resolved.

Beyond improving the generation of URIs, staff should consider further analysis of the different search types. While all three of the search types had considerable usage, the simple search type was associated with nearly 10% more of the log actions than the advanced search type (38.96% for simple versus 29.91% for advanced). Because the simple search type has so much use, it warrants additional attention and development. Users of the search system likely prefer its simplistic format to the more complex set of search parameters associated with the advanced search. More research should be done to ensure that the simple search type, while highly used, also retrieves records effectively and promotes productive search sessions.

Finally, the usage of the advanced search feature suggests that some of the search categories should be reevaluated or reconstituted. Whether a separate credit line category is necessary for a public search system should be considered, especially when a simple keyword search can bring up the same results as the credit line search category. It is clear from the low number of unique queries per session that include this category (0.86%) that very few searchers utilize this field. If this field were removed all users could still use the simple search type to find works related to a particular donor or fund through a keyword search and museum staff could also use TMS as an additional way to access this information. Another option would be to add query completion to this category to see if it

becomes more widely used when individuals are not expected to know specific terms that would be valid for this search parameter.

Limitations

While this study has been able to add to the literature on how individuals seek art and art historical information, the conclusions that can be drawn from it are limited by several factors. Beyond the constraints of the system, the research method selected also limited the nature of the findings. Log analysis, while advantageous for the authenticity and scale of the data it provides, does not give the same depth of information that interviews or other more qualitative methods might. It is ideal to combine log analysis with other research methods, but this was not possible for this study because of time constraints (Peters, 1993, p. 54). Due to the method selected, it was also not possible to define who the users of the search system are. Therefore the conclusions drawn from this study cannot be applied specifically to art historians, artists, or humanities scholars, though individuals who use the site may have fit into one or more of these categories.

One particular research question that developed from this study that cannot be answered through logged actions but could potentially be addressed through more direct contact with users through think-alouds, interviews, or surveys is how the general museum user perceives the department search category. As the most widely used search category in the advanced search type for this study, it would be beneficial to gain a deeper understanding of it. That categories specific to art history, such as department, have not been studied in previous research also justifies focusing on them. In addition to its presence in the advanced search feature on the Ackland's collection search system, the department category is found on many other museum search systems because collections

are divided among separate curatorial departments in order to be managed properly. The department field is also a standard category in TMS, which is one of the most widely used collection management systems today. Because it is so prevalent, seeing if users who are not art professionals accept the different types of sub-categories present within the department category is worthy of research. The department category has fields that relate to culture (e.g. “Asian”), time period (e.g. “Modern and Contemporary”), and form (e.g. “Prints, Drawings, and Photographs”) all under a single heading. While curatorial departments do have a strong influence over the physical organization of objects and exhibits within museums, research to determine whether this category is relevant to the general museum user should be conducted. The high usage of the department category in this study suggests that individuals may not be confused by the conglomeration of sub-categories found within this broader category, but it is important to confirm that this usage actually directly reflects accessibility and that the department category is a meaningful way of organizing the collection for most users.

Conducting this study as a log analysis also made it so that no conclusion could be drawn on the type of information users were seeking. On an art museum collection search system, users could be looking for records in order to see and download images or get information from viewing an image and its associated record. These two goals embody Fidel’s Object Pole and Data Pole respectively. In a lab study one would be able to see if users decided to save an image to their desktop and also have the chance to ask subjects more about the intentions behind their actions. Such a study would also allow researchers to see how users interact with records on the screen and whether study participants typically click on an image or the record’s title to get the complete record.

This information could be used to see if there is a relationship between what is clicked on and the intended use of the information.

Finally, because the Ackland's search system does not include some of the newest features like CBIR technology or subject searching currently found on some museum websites, it was not possible to truly address the relevance of these search capabilities to users. The corpus of queries generated from this study indicates some interest in color searching, but it is negligible. This could be largely because the structure of the Ackland's search system does not encourage these types of queries. Future studies should aim to determine if these and other developing technologies address user needs.

Postscript

Art museum collection search systems provide users with a chance to interact with art that they might not otherwise ever see exhibited in a physical museum or featured in a textbook. These venues are necessarily highly selective and therefore are unable to be representative of the variety of creative work in existence. An art historian interviewed in Bakewell, Beeman, and Reese's 1988 study notes, "In art history in the university you see the top...probably less than 1 percent, of the total number of artworks floating around by the artists that are considered...important...your whole idea of the body of work is formed by the top" (p. 17). Despite the typically narrow scope of art history as a discipline, physical archives as well as online collection search systems can provide users with an opportunity to see beyond this top one percent. Through exploring the works of unrecognized artists or marginal works of those within the accepted canon in either of these contexts, one can "come face-to-face with another kind of reality in the realm of art" (Bakewell, Beeman, & Reese, 1988, p. 17). Especially in cases where the

museum collection exponentially exceeds the exhibition space, like at the Ackland, it becomes increasingly important for the works not on view to be accessible to the public through an alternative method. Online collection search systems have the potential to fulfill this need. If they are built in accordance with user needs, they will likely perform this function well. This is a harder task than it might seem though, because these needs still require further definition. This study has attempted provide some insight into what these needs are.

Notes

¹³ Two instances of links on referring sites that do not resolve include a link to Thomas Eakins works on Wikipedia (https://en.wikipedia.org/wiki/List_of_works_by_Thomas_Eakins) and a link to *The Sword of Damocles* on Learn NC (<http://www.learnnc.org/lp/pages/3059?ref=search>).

REFERENCES

- Bakewell, E., Beeman, W. O., & Reese, C. M. (1988). *Object, image, inquiry: The art historian at work*. Santa Monica, CA: AHIP.
- Bates, M. J., Wilde, D. N., & Siegfried, S. (1993). An analysis of search terminology used by humanities scholars: The Getty online searching project report number 1. *The Library Quarterly: Information, Community, Policy*, 63(1), 1-39.
- Beard, C. L. (2004). *An Internet search interface for the Ackland Art Museum Collection Database* (Unpublished Master's paper). University of North Carolina, Chapel Hill.
- Belkin, N. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, 5, 133-143.
- Cameron, F. (2012). Museum collections, documentation, and shifting knowledge paradigms. In G. Anderson (Ed.), *Reinventing the museum: The evolving conversation on the paradigm shift* (223-238). New York: Altamira Press.
- Chau, M., Fang, X., & Sheng, O.R.L. (2005). Analysis of the query logs of a web site search engine. *Journal of the American Society for Information Science and Technology*, 56(13), 1363-1376.
- Chen, H.-L. (2001). An analysis of image queries in the field of art history. *Journal of the American Society for Information Science and Technology*, 52(3), 260-273.
- Choi, Y. & Rasmussen, E. M. (2003). Searching for images: The analysis of users' queries for image retrieval in American history. *Journal of the American Society for Information Science and Technology*, 54(6), 498-511.
- Cobbledick, S. (1996). The information-seeking behavior of artists: Exploratory interviews. *The Library Quarterly: Information, Community, Policy*, 66(4), 343-372.
- Croft, W.B., Metzler, D., & Strohman, T. (2009). Crawls and feeds. In *Search engines: Information retrieval in practice*. Cambridge: Cambridge University Press.
- Di Grazia Costa, I., Napoleone, L. M., & Da Rocha, V. G. (2012). Transaction log analysis of the use of art information resources at the São Paulo Museum of Art. *Art Libraries Journal*, 37(4), 31-35.

- Dumais, S., Jeffries, R., Russell, D. M., Tang, D., & Teevan, J. (2014). Understanding user behavior through log data and analysis. In J.S. Olson and W. Kellogg (Eds.), *Human Computer Interaction Ways of Knowing*, 349-372. New York: Springer.
- Enser, P. G. B. (1993). Query analysis in a visual information retrieval context. *Journal of Document and Text Management*, 1(1), 25-52 .
- Fidel, R. (1994). User-centered indexing. *Journal of the American Society for Information Science*, 45(8), 572-576.
- Fidel, R. (1997). Image retrieval task: Implications for the design and evaluation of image databases. *The New Review of Hypermedia and Multimedia*, 3, 181-199.
- Gayo-Avello, D. (2009). A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, 179, 1822-1843.
- Goodrum, A. & Spink, A. (2001). Image searching on the Excite Web search engine. *Information Processing and Management*, 37, 295-311.
- Gregory, T. R. (2007). Under-served or under-surveyed: The information needs of studio art faculty in the southwestern United States. *Art Documentation*, 26(2), 57-66.
- Hare, J. S., Lewis, P. H., Enser, P. G. B., & Sandom, C. J. (2006). Mind the gap: Another look at the problem of the semantic gap in image retrieval. In *Multimedia Content Analysis, Management, and Retrieval 2006*, 17-19 January, San Jose, California, USA. doi:10.1117/12.647755.
- Hastings, S. K. (1994). *An exploratory study of intellectual access to digitized art images* (Unpublished doctoral dissertation). The Florida State University, Tallahassee.
- Hekkert, P. & Van Wieringen, C. W. (1996a). Beauty in the eye of expert and nonexpert beholders: A study in the appraisal of art. *The American Journal of Psychology*, 109(3), 389-407.
- Hekkert, P. & Van Wieringen, C. W. (1996b). The impact of level of expertise on the evaluation of original and altered versions of post-impressionistic paintings. *Acta Psychologica*, 94(2), 117-131.
- Hemmig, W. S. (2008). The information-seeking behavior of visual artists: A literature review. *Journal of Documentation*, 64(3), 353-362.

- Isemann, D. & Ahmad, K. (2011). Query terms for art images: A comparison of specialist and layperson terminology. In *Proceedings of the 25th BCS Conference on Human-Computer Interaction* (BCS-HCI '11). British Computer Society, Swinton, UK, 145-150.
- Jansen, B. J. (2006). Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research*, 28, 407-432.
- Jørgensen, C. (1995). Image attributes: An investigation (Unpublished doctoral dissertation) Syracuse University, Syracuse.
- Jørgensen, C. (1996). Indexing images: Testing an image description template. *Proceedings of the American Society for Information Science Annual Meeting*, 33, 209-216.
- Jørgensen, C. (1999). Retrieving the unretrievable in electronic imaging systems: Emotions, themes, and stories. Paper presented at Human Vision and Electronic Imaging IV, *SPIE* 3644, 348-355. <http://dx.doi.org/10.1117/12.348455>
- Jørgensen, C. (2003). *Image retrieval: Theory and research*. Lanham, MD: The Scarecrow Press, Inc.
- Liu, Y., Zhang, D., Lu, G., & Ma, W.-Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40, 262-282.
- Lowe, H. A. (2013). Search log analysis of the ARTstor Cultural Heritage Image Database (Unpublished Master's paper). University of California at Los Angeles, Los Angeles.
- Lunin, L. F. (1994). Analyzing art objects for an image database. In R. Fidel, R. B. Hahn, E. M. Rasmussen, & P. J. Smith (Eds.), *Challenges in indexing electronic text and images* (57-72). Medford, NJ: Learned Information.
- Markey, K. (2007). Twenty-five years of end-user searching, part 1: Research findings. *Journal of the American Society for Information Science and Technology*, 58(8), 1071-1081.
- Marty, P. F. (2007). Museum websites and museum visitors: Before and after the museum visit. *Museum Management and Curatorship*, 22(4), 337-360.
- Marty, P. F. (2008). Museum websites and museum visitors: Digital museum resources and their use. *Museum Management and Curatorship*, 23(1), 81-99.

- McLaughlin, M., Goldberg, S. B., Ellison, N. & Lucas, J. (1999). Measuring Internet audiences: Patrons of an online art museum. In S. Jones (Ed.), *Doing Internet research: Critical issues and methods for examining the net* (163-178). Thousand Oaks: Sage Publications, Inc.
- Mitroff, Dana and Katrina Alcorn. (2007). Do you know who your users are? The role of research in redesigning sfmoma.org. In D. Bearman and J. Trant (Eds.), *Museums and the Web 2004* (11-20). Toronto, Canada: Archives & Museum Informatics. Retrieved from <http://www.archimuse.com/mw2007/papers/mitroff/mitroff.html>
- Peacock, D., Ellis, D., & Doolan, J. (2004). Searching for meaning: Not just records. In D. Bearman and J. Trant (Eds.), *Museums and the Web 2004* (11-20). Toronto, Canada: Archives & Museum Informatics.
- Peters, T. A. (1993). The history and development of transaction log analysis. *Library Hi Tech*, 11(2), 41-67.
- Rieh, S. Y. & Xie, H. I. (2006). Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing and Management*, 42, 751-768.
- Shatford, S. (1986). Analyzing the subject of a picture: A theoretical approach. *Cataloging & Classification Quarterly*, 6(3), 39-62. http://dx.doi.org/10.1300/J104v06n03_04
- Skov, M. & Ingwersen, P. (2014). Museum web search behavior of special interest visitors. *Library & Information Science Research*, 36, 91-98.
- Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 22(12), 1349-1380. <http://dx.doi.org/10.1109/34.895972>
- Taghavi, M., Patel, H., Schmidt, N. Wills, C., & Tew, Y. (2012). An analysis of web proxy logs with query distribution pattern approach for search engines. *Computer Standards & Interfaces*, 34, 162-170.
- Taylor, R. S. (1968). Question-negotiation and information seeking in libraries. *College & Research Libraries*, 29(3), 178-194.
- Wells-Angerer, T. L. (2005). *A study of retrieval success with original works of art comparing the subject index terms provided by experts in art museums with those provided by novice and intermediate indexers* (Unpublished Master's paper). University of North Carolina, Chapel Hill.

Wiberley, S. E., Jr. (1988). Names in space and time: The indexing vocabulary of the humanities." *Library Quarterly*, 58, 1-28.

Wildemuth, B. M. (2009). *Applications of social research methods to questions in information and library science*. Westport, CT: Libraries Unlimited.

ART CAPTIONS

Albert Bierstadt, American, 1830-1902: Blue Mountain and Lake, 1857-1862, oil on paper, mounted on board, 10 1/2 x 15 1/2 in. (26.6 x 39.3 cm) Ackland Art Museum, The University of North Carolina. Gift of Charles Tate, 94.14

Agnolo Bronzino, Italian, Florence, 1503-1572: The Virgin and Child with the Infant St. John the Baptist, 1560s, oil on wood pane, 134 1/16 x 26 7/8 in. (86.5 x 68.3 cm) Ackland Art Museum, The University of North Carolina. Ackland Fund, 64.28.1

Attributed to Bucci Painter, Greek, Attic, 6th century B.C.: Vessel (Neck Amphora) with Apollo, Leto, and Artemis, c. 540-530 BCE, terra cotta, black-figure ware, 15 15/16 x 11 7/16 in. (40.5 x 29.1 cm) Ackland Art Museum, The University of North Carolina. Ackland Fund, 88.15

Circle of Lucas Cranach the elder, German, 1472-1553: The Mass of St. Gregory, c. 1550, oil on wood panel, 34 x 24 3/8 in. (86.4 x 61.9 cm) Ackland Art Museum, The University of North Carolina. Ackland Fund, 59.8.2

Edgar Degas, French, 1834-1917: Spanish Dance, c. 1885, cast 1921, Bronze, 18 1/4 x 5 5/8 x 8 3/4 in. (46.3 x 14.3 x 22.2 cm) Ackland Art Museum, The University of North Carolina. Ackland Fund, 74.21.1

Eugène Delacroix, French, 1798-1863: Cleopatra and the Peasant, 1838, oil on canvas, 38 1/2 x 50 in. (97.8 x 127 cm) Ackland Art Museum, The University of North Carolina. Ackland Fund, 59.15.1

Robert Freebairn, British, 1764-1808: Falls of Tivoli, 1807, Oil on canvas, 36 x 48 in. (91.4 x 121.9 cm) Ackland Art Museum, The University of North Carolina. From the Ruth and Sherman Lee Collection, Gift of Katharine Lee Reid, 2012.41

Nicolas Lancret, French, 1690-1743: Dance in a Garden, mid-1730s, Oil on canvas, 23 7/16 x 20 in. (59.5 x 50.8 cm) Ackland Art Museum, The University of North Carolina. Ackland Fund, 72.22.1

Master of 1419, Italian, Florence, active c. 1419-1430: The Virgin and Child, c. 1415, tempera and gold on wood pane, 145 13/16 x 21 3/8 in. (116.4 x 54.3 cm) Ackland Art Museum, The University of North Carolina. Ackland Fund, 80.34.1

Camille Pissarro, French, 1830-1903: The Banks of the Oise, Near Pontoise, 1876, oil on canvas, 14 15/16 x 21 7/8 in. (38 x 55.5 cm) Ackland Art Museum, The University of North Carolina. Ackland Fund, 65.28.1

Pieter Symonsz Potter, Dutch, c. 1597-1652: Lot and His Family Fleeing from Sodom, n.d., oil on panel, 23 1/8 x 18 1/2 in. (58.7 x 47 cm) Ackland Art Museum, The University of North Carolina. Bequest of Henry W. Lewis, 2005.4.2

Unidentified Artist: Egyptian Princess with a Musical Instrument, c. 1360-1350 BCE, white sandstone, 6 1/4 x 4 1/2 x 1 1/4 in. (15.9 x 11.4 x 3.2 cm) Ackland Art Museum, The University of North Carolina. Ackland Fund, 67.29.4

Unidentified Artist: Seated Girl Holding a Blossom, late 17th century, ink and watercolor, 3 15/16 x 3 1/16 in. (10 x 7.8 cm) Ackland Art Museum, The University of North Carolina. Ackland Fund, 85.16.1

Unidentified Artist: St. Jerome in Penitence, c. 1515, oil on wood pane, 145 5/8 x 33 1/2 in. (115.9 x 85.1 cm) Ackland Art Museum, The University of North Carolina. The William A. Whitaker Foundation Art Fund, purchased in memory of Clemens Sommer, Professor of Art 1940-1962, 67.31.1

Richard Westall, British, 1765-1836: The Sword of Damocles, 1812, oil on canvas, 51 3/16 x 40 9/16 x 3 1/4 in. (130 x 103 x 8.3 cm) Ackland Art Museum, The University of North Carolina. Ackland Fund, 79.10.1