User Access to Digital Image Collections of Cultural Heritage Materials: The Thesaurus as Pass-Key

Johanna Woll

[The following article was the winner of the 2005 Gerd Muehsam Award. The award recognizes excellence in a paper written by a graduate student on a topic relevant to art librarianship or visual resources curatorship.]

Introduction

Technological trends of the past ten years have made it possible to offer users a range of library resources in a variety of formats through searchable online databases. Optical character recognition, a variety of image file formats, scanning technologies, storage media, and high-resolution digital capture devices have all become less expensive and more sophisticated and flexible. These changes are a boon to visual resources collections, especially at a time when equipment for analog capture and presentation are at the cusp of a downward spiral toward obsolescence.[1] At the same time, visual resources professionals now face myriad and previously unencountered challenges as they strive to convert their analog collections to new formats: selecting images for conversion, designing new workflows, developing instruction modules for patrons and training for staff, and installing new software and hardware needed for access, delivery, and presentation. Furthermore, providing intellectual access to digital images focuses attention even more closely on the paramount need for effective description, indexing, and retrieval.

In this article I explore research relating to thesaurus-assisted searching and user access to digital image collections of cultural heritage materials,[2] in particular subject access, and to the potential for this type of partially mediated searching, using a deep semantic network, to improve intellectual access to digital image collections. Many studies have investigated various components of this topic, though primarily in the context of textual resources; fewer have looked at the intersection of these components in the context of visual resources. There are many complex interrelationships among user types, behaviors, and needs, thesaurus types and designs, subject access, search systems, interface design, and system performance, and they all affect the indexing and retrieval of digital images of cultural heritage materials.

First, I must articulate some of the assumptions, supported by research findings and analysis in information retrieval, indexing, and user behavior, which form the framework for this discussion. These include:

- Indexing and end-user retrieval are closely linked; better indexing translates to better retrieval
- Indexers are likely to use different subject terms from those used by searchers; searchers tend to use natural language terms in their queries rather than terms selected from an authoritative vocabulary or list of subject headings
- Certain patterns emerge from an analysis of image queries (e.g., frequently used query terms refer to creator name, and generic and unique objects and events)
- Image titles or captions may not effectively convey subject content
- Users, even highly educated and intelligent scholars, are often not skilled searchers of online databases; they can be resistant to learning and applying relatively simple search operators and command languages; we must, therefore, seek to develop search tools that require minimal additional effort on the part of users

Definitions

I use the term images to refer to digital files or analog documents that function as surrogates for original works of art (including art photographs), the built environment, maps, and other cultural heritage objects; that is, the kinds of still images typically found in a visual resource collection serving the needs of patrons in an academic environment.

I define thesaurus as a structured vocabulary designed to ensure authority control for terms used in descriptive cataloging and to facilitate effective retrieval for users searching an online database. Thesauri generally consist of postcoordinated single terms focusing on a specific domain of knowledge. Many are "faceted" or structured around an internal classification system. I do not use thesaurus to refer to the type of general reference work that is used to identify synonymous words, a Roget's-type thesaurus.

The Getty Research Institute has developed thesauri specifically to serve indexers and researchers working with cultural heritage materials. These include the *Art and Architecture Thesaurus (AAT)*, the *Thesaurus of Geographic Terms (TGN)*, and the *Union List of Artist Names (ULAN)*. In the 1940s Henri van der Waal at the University of Leiden created *ICONCLASS*, a hierarchically-structured index for iconographical content. The Library of Congress has built the *Thesaurus for Graphic Materials I* and *II (TGMI* and *TGMII)*.

Library of Congress Subject Headings (*LCSH*) may also be applied to cultural heritage materials, but this list of authorized headings is not a thesaurus in the strictest sense of the term.

While offering references to broader, narrower, related, and equivalent terms as thesauri do, *LCSH*, as an alphabetical list, is not formally structured hierarchically, syndetically, or categorically (using facets). Many visual resource collections have modified existing thesauri or developed local thesauri to meet the particular needs of their collections.

Peculiarities of Cultural Heritage Materials

Cultural heritage materials offer researchers information-rich resources that can be used in diverse contexts. A single image of a wall painting from an eighth century palace in Jordan may help an architectural historian identify the patron of the building complex, a conservator identify early materials and techniques, a sociologist identify status symbols and class structures, and a medical historian identify personal hygiene customs. Because of its dense information content and its appeal to diverse users, this same image presents unique challenges for cataloging and searching.

The very act of categorizing an image as a cultural heritage resource is problematic, for an image that represents simple documentation to one viewer and is sought for its "evidential value" may be of great artistic or cultural relevance to another who seeks its "value as an artefact."[3] Over time, an image's designation may very well change: it may come to represent something that has since disappeared or something that has become interesting or significant where once it was merely banal.

Peculiarities of Intellectual Access to Images

Treating images, and in particular digital images, as one would treat textual materials has proven to be a losing strategy; information specialists dealing with images have therefore worked to develop different metadata schemas, different subject vocabularies, and different cataloging rules. Digital images amplify their differences from textual materials even more. Whereas access to analog image collections depends on the physical order of the materials (which are usually not arranged by subject), access to digital image collections depends on matching users' queries to indexing terms derived from both bibliographic-type data and subject terms. Online access and retrieval is at once more flexible and more difficult, more intuitive and less logical. In an analog environment, visual cues like the age of a slide mount, the type of label, the typeface, and the quality of the image all provide additional information about an image at hand. None of this is immediately observable with digital images. With no logical shelf order, the process of browsing is different, as is the capacity for serendipitous discovery. A digital collection may be coherent from a curatorial perspective, but with its images distributed arbitrarily in a database, it is diffuse from a viewer's perspective.

Images bear their meaning not in natural language, but in the arrangement of their colors, shapes, textures, dimensions—in short, their physical and visual attributes. Viewers perceive and interpret images in dramatically different ways, and this diversity necessarily leads to subjectivity in the choice of terms used to describe images. Indeed, it may be impossible to describe an image objectively. This intrinsic dissonance makes harmonizing the two poles of access—indexing and searching—exceptionally difficult. In all contexts, users are most likely to retrieve satisfactory results when they understand the structure and language of the search system and of the indexing terms used.

The literatures of education, art history, psychology, and cognitive science have all described how visual information differs from verbal information. Shatford notes that the "subjects of pictures have essential qualities that make them different from the subjects of textual works."[4] Jörgensen notes that access points for images are necessarily more numerous because of the wide variety of information that they contain.[5] Chen and Rasmussen refer to "hard" and "soft" indexing,[6] the former a description of the image's visual features, the latter its meaning. Searchers seek both types of image attributes.

Applying subject terms to images differs from applying subject terms to textual resources primarily because with images there are few words, and in some instances no words, that describe what an image depicts, what it is about, what it means. Books and other textual materials offer indexers and descriptive catalogers titles and subtitles, abstracts, tables of contents, chapter titles, and of course the actual contents that directly or indirectly indicate the subject of the work. Image titles and captions may not offer the same guidance (consider, for example, a painting titled *Abstract Composition* or *Untitled*).

An image surrogate is a distinct entity separate in time and space from the original work of art, and some of the surrogate's attributes — for example, dimensions or creation date — are likely to differ from those of the original. Although the same may be true of various editions of a given textual work, there may be greater inherent ambiguity for a visual work. Many searchers may not be aware of the distinction between a work (the original) and its surrogates (any number of reproductions of that original). Shatford highlights these distinctions in her discussion of Represented Work,[7] that is, a work appearing within another work (a painting of a cathedral, for example). Furthermore, multiple image surrogates may be associated with the same original work; for example, different views of a single sculpture or building.

Online image searching is largely unmediated, with no reference interview to help clarify terms and define search parameters. Cawkell suggests "since there may be a multiplicity of entities within a single image, and a multiplicity of attributes associated with any one of the entities, it is contended that the subject requests for which a given image may be deemed relevant, appropriate, or interesting, are much less predictable than is the case for textual material."[8]

Translating concepts and visual perceptions to words, especially to specialized terms that might appear in an index or thesaurus, and mapping the terms of a mental model or natural-language query to controlled language, are fundamentally different from the processes of subject cataloging of texts. Indeed, Svenonius, writing about the problem of translating subject concepts from one medium to another as this relates to indexing nonbook materials, concludes, "there are instances that defy subject indexing."[9]

Svenonius cites Langer, who asserts that music and works of art convey "presentational symbolism" which in no way resembles verbal language. While Svenonius disagrees (countering that visual language does have a kind of lexicon),[10] she concedes that "even if we could construct lexicons of visual forms and musical notes, they would be logically different in kind from word lexicons. Proof of this is that visual forms and musical notes, unlike words, can neither be defined nor translated."[11] So presented with this dilemma, how does one apply verbal subject terms to visual materials?

Thesauri

Thesauri were originally developed in the 1950s when they primarily addressed the needs of indexers and searchers in the fields of science and technology. In time, their topical coverage extended to the social sciences and humanities.[12] Thesauri represent subjects in terms of their hierarchical (broader, narrower), associative (related), and equivalence (synonyms, variants) relationships. Thesauri are different from controlled lists and other types of authority sources that allow users only to select terms, not to perceive relationships among terms.

The purpose of a thesaurus, according to Bearman, is to suggest "more precise terms to indexers and users," to define "the relationships among terms (broader, narrower, synonymous)," and "to identify materials indexed by a term different than the one [the users] know."[13] In indexing, the framework created by a thesaurus' inherent hierarchy facilitates the selection of appropriate terms, and because thesauri cluster variants for a single term, searchers may use any one of the variants as a query term; the search system will look for matches on all of the variants. Scope notes are sometimes included for terms that might be ambiguous. Most thesauri, while allowing for the construction of strings or expressions, consist of postcoordinated single terms, unlike *LCSH* and other lists that include multifaceted subjects rather than specific concepts.[14]

Greenberg lists characteristics of the *AAT* and of the *TGM* which are also common to other thesauri: "Single terms are used for concrete concepts; plural terms are used for abstract ideas; gerunds are used for activities; parenthetical qualifiers are used where necessary;[15] compound terms are used when a single term cannot fairly serve; dated terms are eliminated, while new terms are added; user [indexer] participation is encouraged."[16]

Thesauri must strike a balance between specificity and flexibility. Terms that are too specific may be used by few indexers and searchers. Yet a thesaurus is more flexible if specific terms are included and accessible to a search system that might expand a query by automatically retrieving broader or narrower terms.

Faceted Vocabularies

Hierarchically-structured thesauri often consist of facets which, as single concepts by definition, avoid the reported weaknesses of precoordination while still supporting postcoordination at the time of a search.[17] Facets, as defined by Petersen, are "homogeneous, mutually exclusive units of information which share characteristics that demonstrate their differences from each other."[18] The theory of facets was pioneered in the 1930s by British classification theorist S.R. Ranganathan whose five categories—Personality, Matter, Energy, Time, and Space — would become the basis for many subsequent faceted vocabularies.[19]

Bates asserts that although "faceting provides a flexibility and power much in excess of the limited character of conventional hierarchical schemes,"[20] many information systems continue to use conventional ("family-tree style") classification schemes. Using faceted thesauri that cluster term variants arguably improves intellectual access and simplifies the process of indexing.[21] A Getty Information Institute study of humanities scholars found that 91 percent of the searches performed included subject terms associated with distinct categories—works or publications, individuals, geographical, chronological, discipline—categories similar to those in faceted thesauri.[22]

Mills identifies faceted classification as the "only viable form enabling the locating and relating of information to be predictable…to understand anything, whether it is the operation of a complicated mechanism or the complex social factors that underlie almost any human situation, understanding it means seeing the connections."[23] Mills argues that subject-heading lists not only lack the specificity needed for special collections (like visual resources), but also suffer from the "relative arbitrariness in the provision made for the relating function."[24]

O'Neill and Chan describe an OCLC research project called *FAST (Faceted Application of Subject Terminology)*, a proposed controlled vocabulary based on *LCSH* that would address many of the shortcomings of *LCSH* and other subject-heading lists and better meet the needs of online searchers. Like *LCSH*, terms would be added on the basis of literary warrant.[25] Unlike *LCSH*, *FAST* would be structured using facets.

AAT's facets are eminently suited to visual resources of cultural heritage materials as are the *ICONCLASS* facets, though the latter suffers from a complicated and onerous notational system, a bias toward the Western world, and a lack of quotidian descriptive terms.[26]

Hierarchical Relationships

Many thesauri possess a hierarchical structure with explicit superordinate and subordinate relationships; for example, genus-species, part-whole, parent-child, or topic-subtopic. To designate these relationships, thesauri employ standard codes—broader term (BT) and narrower term (NT)—although these designations may take on subtly different meanings in different contexts.[27]

ICONCLASS, for example, combines a faceted classification scheme, designed with the same type of hierarchical relationships found in other thesauri, with a kind of subject-heading list. *ICONCLASS* links its preferred term (single word or multiword phrases) to related keywords. These latter are not necessarily hierarchically related terms; rather, they are different keywords that lead users to an authorized subject term.[28]

Associative Relationships

Associative relationships include those that consist of cross references to other terms. Although some controlled vocabularies have no explicit hierarchy and are arranged alphabetically, the syndetic terms of associative relationships (references primarily to BTs, NTs, but also to RTs and similar cross references) offer an implicit hierarchy. Related terms can be problematic, however, because they can designate numerous types of relationships: associated fields of study, causal dependence, or actions and outcomes. Moreover, associative relationships in the form of RTs may refer either to terms in the same facet or to terms from entirely different facets.

Over the years, *LCSH* has slowly been eliminating its "see" and "see also" cross references, replacing these with the more common thesaural codes of BT, NT, and RT. *LCSH* retains, however, its "use for/see from" references and strives to minimize the use of related terms.[29] In *AAT*, "use for" terms are treated as variant or equivalent terms.

Equivalence Relationships

Most terms possess at least one morphologically or syntactically close variant. For example, there are singular or plural

forms, or inverted forms of modified or compound terms. Synonyms, variant spellings, foreign-language terms, chronological variants (developed as preferred usage changes over time) also represent equivalence relationships. In many cases, equivalence relationships in thesauri eliminate the need for long strings of Boolean OR statements, which searchers have traditionally used to ensure retrieval of variant forms of a term. For digital image databases, where there is little or no text to scan in a keyword search, equivalence relationships may accommodate natural-language, keyword-type searches in cases where the natural-language terms are listed as variants.

Structured vocabularies ensure authority control, facilitate the modification of queries, and largely shift the burden of selecting an appropriate term from the (untrained) searcher to the trained indexer or to the search system itself. Faceted vocabularies may also serve as a classification scheme, categorizing terms into similar groupings. In an online environment, where documents lack the intuitive cues of collocation, this characteristic may help users make sense of retrieved results.

In spite of their many advantages, thesauri and other vocabularies must not be considered a panacea for online searchers. Inconsistencies and ambiguities still detract from their effectiveness. For example, many vocabularies have neither articulated guidelines nor defined the intellectual nature of terms to be used, what Rolland-Thomas calls the "level of discourse."[30] Should the preferred term be the common, natural-language term that is most likely to pop into the searcher's head, or should it be the technical, domain-specific, or expert term?

Furthermore, different systems apply different definitions to thesaural codes (BT, NT, RT); indeed, hierarchical terms may imply different types of relationships in different disciplines. More research into the development of common definitions and explicit guidelines for assigning these terms, standards that accommodate all knowledge domains, would ultimately improve intellectual access.[31] O'Neill and Chan articulate the key requirements for any subject vocabulary to be used in an online environment: simple in structure and easy to maintain; able to provide optimal access points; flexible and interoperable across disciplines.[32] They call for clear definition of the semantics of the terms (choice of vocabulary) and of their syntax.

Searchers: Myriad Types, Needs, and Behaviors

The diversity of users of online search systems reflects the great diversity of humankind, with its multitude of attitudes, intellectual skills, and cognitive styles. Some searchers have deep subject-matter expertise and conduct known-item searches. Some have much online searching experience and are familiar with Boolean operators, truncation, and other online search strategies. Others possess neither domain expertise nor search skills.

Users of digital image collections demonstrate different needs and use information in images differently. Images might be compared, analyzed individually or in groups, manipulated, presented out of context, or used to illustrate a very specific point. Brilliant describes art historians as scholars whose paradigm pivots on the connection of art objects to their historical and contextual significance.[33] Roberts reports that visual resource collections are increasingly used by more types of patrons, and that traditional patrons (primarily art historians) are using the images in nontraditional ways, analyzing images not only for their aesthetic and stylistic characteristics but also for their value as historical documents with iconographical, political, theological, and ideological significance.[34] These changes call for a revision of our conception of subject access and increased efforts to serve a broader range of user types.

A study conducted by Bates and others at the Getty Information Institute revealed at the time (1988-90) that the search terms used by humanities scholars differ significantly from those available in conventional thesauri. Bates's research challenges the commonly held belief that humanities search terms are characteristically less precise than those used in the sciences. A National Science Foundation study of social science subject-term queries found that 100 percent of queries contained at least one common (uncapitalized) term. In contrast, only 57 percent of the humanities queries in the Getty Information Institute study of humanities scholars had at least one common term.[35] Perhaps not surprisingly then, many thesauri emphasize common terms; thesauri for the humanities will need to emphasize other types of terms as well (e.g., proper nouns), and indeed, the Getty's *ULAN*, *AAT*, and *TGN* have addressed this need.

Humanities scholars also demonstrate a wide variety of preferences in their search strategies (mediated, partially mediated, unmediated, etc.). Scholars in the Getty study found Boolean search operators and the "logical, engineer-oriented design of online systems"[36] ill-suited to their native mode of constructing queries.

Shatford Layne describes research by R.M. Diamond conducted in 1969 in which faculty members from the disciplines of history, literature, and art history analyzed a set of slides and selected terms that they would want to use to access them. Each discipline focused on different attributes of the images.[37] Another study, by Sutherland, also demonstrated the extent to which users interpret images differently; in this case, an art historian, a layperson, and a twelve-year-old child described the same images and, not surprisingly, each used different types of terms and found different visual and iconographic elements.[38] Enser and McGregor, examining queries posed by searchers of a large general picture collection (analog, not digital), identified four categories of query type: unique, nonunique, refined, nonrefined.[39]

Hastings conducted a small-scale study of art historians accessing a digital image collection of Caribbean paintings at the University of Central Florida. She sought to understand if search behavior for digital image collections mirrored or differed from that for analog collections and whether additional indexing elements might be required to provide satisfactory access. Four query categories of incremental complexity emerged. Hastings concluded that art historians do demonstrate a particular approach to searching online collections, one that is significantly more complex than that used to search analog collections (for example, browsing digital thumbnails in a refinement phase of the search and modifying queries using search functions).[40]

Subject Naming

Much research has demonstrated that subject terms represent a high percentage of search terms used (nevertheless, art historians, regardless of their approach to accessing image collections, also seek basic descriptive data like title, creator, date, location, and style and period).[41] Good subject access relies

on accurate indexing with appropriate subject terms, a process that involves two steps: analysis and identification of concepts, and translation of these into appropriate verbal terms. Indexers, according to Dykstra,[42] must match their level of analysis with the level of detail available in the target vocabulary.

Erwin Panofsky proposed three levels of subject analysis that cover all aspects of an artwork's content: pre-iconographic description (generic elements), iconographic identification (named elements), and iconographic interpretation (meanings or themes).[43] Shatford modifies Panofsky's levels by defining additional categories that she calls Generic Of, Specific Of, and About.[44]

Of-ness is articulated with concrete terms that describe what an image depicts through either generic or specific (named) objects, events, places, and conditions. About-ness refers to abstract concepts, moods or emotions, mythical beings, and symbolic elements that are derived from analysis and extend beyond mere description. Both of these analytical categories may include attributes relating to form, genre, style, or date, often considered separate data elements in a catalog record. Search systems may need to conflate some of these attributes into the subject term to accommodate query terms; indeed, in many cases, the terms for form, genre, and style are selected from the same controlled source as the subject terms.

Because different searchers derive different kinds of value from an image, each level of subject analysis must be recorded for optimal access. Known-item searching may only require pre-iconographic or descriptive elements, whereas more complex thematic searches may require iconographic terms. Shatford Layne argues that user queries seek both of-ness and about-ness content, with of-ness a relevant factor for 35 percent of art-related research and about-ness relevant for 20 percent of art-related research.[45]

Brown's research suggests that subject naming of complex concepts, like those commonly found in images, is correlated to the concreteness, complexity, and syndeticity of the query terms used. She argues that because subject searching predominates in online catalogs and because matching query terms to indexed terms has a low likelihood of success (with correct subject headings used in queries only 20-35 percent of the time), we must focus our attention on improving subject naming by prompting users to select simple (not compound), concrete (not abstract) terms, and to refine their searches using related concepts rather than different hierarchical levels (superordinate or subordinate).[46]

Query Construction in Online Searching

The way that searchers construct their queries using an online system depends in part on the system interface. Many search interfaces for textual resources (e.g., OPACs and commercial databases) offer users query categories, in effect prompting them to specify the associated metadata element (or database field) to be searched. Digital image collections often offer the same feature (Library of Congress Prints and Photographs Online Catalog, University of California, Berkeley's SPIRO), but some do not (the Gertrude Bell Archive, David Rumsey Map Collection). Some systems, while accommodating natural-language terms, also offer a browsable thesaurus to help users identify authorized terms (*ARTbibliographies Modern, ArtAbstracts*), while others have no browsing function and list authorized terms only in the citations of retrieved documents (*Bibliography of the History of Art, Avery Index to Architectural Periodicals*).

For online catalogs, shelf order, collocation, and the problem of "distributed relatives" created by the use of precoordinated terms are largely irrelevant. Access points are treated as equal in importance; searchers may enter at any level, through any data element. Searchers may select multiple constituent data elements or subject terms rather than having to identify (or even know) the main entry in the traditional sense of the catalog record. However, this flexibility can lead to increased ambiguity. Many search interfaces do not offer qualifiers for search terms and, without proper precoordination, many search strings will yield unsatisfactory, albeit legitimate, results.

Bates proposes a design model for online subject access that includes an end-user thesaurus and a "front-end system mind," arguing that such a model would respond to the recent shift that favors subject access over access through descriptive elements.[47] She cites a Council on Library Resources study that found that users accessing online catalogs used subject terms 59 percent of the time.[48] Based on the principles of uncertainty, variety, and complexity, Bates's model is constructed around three phases: access (including entry and orientation), hunting, and selection.[49]

The uncertainty principle refers to the impossibility of predicting the terms that an indexer or searcher will use to describe or locate an item. Bates recommends accepting this inherent uncertainty rather than attempting to eliminate it: "Stop trying to design systems that will target the desired information through perfect pinpoint match on the one best term; rather, design systems to encompass the answer by displaying and making it easy to explore a variety of descriptive terms."[50]

The variety principle refers to Ashby's law, which states that the diversity of indexing and query terms must be equal. One can satisfy this requirement by reducing variety at one end and increasing it at the other: reducing indexer variety through the use of controlled vocabularies and reducing search-term variety by providing rich cross references. Thesauri, of course, accommodate both of these requirements.

The complexity principle refers to the inherent complexity of interactions in an information system. To address this characteristic, Bates proposes a front-end system mind (FSM), a rich semantic network constructed from one or many vocabulary sources, which handles the complex process of matching a query to indexed terms. Because the FSM and its network of relationships is visible to the user, as is the thesaurus, this approach transfers the burden of generating appropriate terms and finding related terms from the searcher to the system; searchers are left to focus on what is most important, the selection of materials that match their needs.

Bearman points to a number of unresolved functionality issues that limit the usability of thesauri (specifically the *AAT*) for searching an online collection.[51] He argues that even with systems able to handle many types of terms (uniterms, expressions, constructed strings), users are likely to be unfamiliar with the structure of the thesaurus and the relationships among terms and may not conceive their search optimally. To address this limitation, Bearman proposes several functional requirements.

First, a broad search term must automatically search both the search term itself (and all its variants) and all narrower terms that fall under the broad term in the hierarchical structure of the thesaurus. He calls this "exploding on a term." The same requirement would hold for search expressions.

Second, if the search interface is "fielded" and allows users to enter search terms for specific facets or categories, a null value in a given search field will necessarily broaden the search to include all possible values for the blank fields. However, users must be made aware of this default so that they may further refine their searches to eliminate unwanted categories of results.

Third, because equivalent or related terms may appear under more than one facet, when using an expression (often a modified term), users may inadvertently broaden their search quite dramatically, since each term in the expression may sit under several hierarchical levels or different facets. Bearman suggests that a web of links in the thesaurus might accommodate these overlaps and relationships but admits that implementing this feature may be exceedingly difficult. Like Bates, he suggests making the hierarchical structure visible to users so that they can better understand the relationships of their search terms to other terms in the thesaurus and can adjust their searches accordingly. This might best be accomplished, according to Bearman, with a mouse-over function that would explode the immediate hierarchy of a given search term, enabling iterative searches in a multiple-window environment and displaying retrieved results for multiterm searches in ranked order.

Query Modification in Online Searching

Query expansion, whether automatic, dynamic, or user-activated, describes the act of reformulating a query after observing the retrieval results, usually by adding or deleting search terms.[52] Greenberg conducted research to determine whether automatically expanding a subject-term query using specific types of thesaurus terms related semantically to the query—NTs, BTs, RTs, synonyms and partial synonyms—might influence the relative performance of retrieval. She found that query expansion using synonyms and narrower terms increased recall with no significant effect on precision. Query expansion using broader and related terms also increased recall but with a significant effect on precision. These results and others cited by Greenberg support the assertion that structured thesauri can contribute to improved retrieval. The automatic nature of some query expansion features responds to the observed difficulty that many users experience in trying to modify their query appropriately using authorized terms of the controlled vocabulary. The inherent knowledge structures of faceted thesauri offer valuable cues for modifying queries and improving the search process.

Shatford Layne discusses an initial search step involving querying subject term(s) and a subsequent step of sorting or clustering retrieved results by using an additional refiner, or qualifier, especially for cases where the same subject term might apply to different descriptive data elements (for example, Goya could be a creator name or Goya could be the subject of a work, but a system might not distinguish between a work *by* Goya and a portrait *of* Goya by another artist).[53]

Dykstra writes that because their query terms are more likely to appear in a variety of syntactical arrangements, searchers from knowledge domains like humanities, the arts, and social sciences may benefit from a browse feature that would allow them to select from retrieved results.[54] Turner and others have advocated a hybrid approach to image searching that combines an initial text search with a visual search of a retrieved set of thumbnail images.[55] Shatford Layne agrees: "Rather than devoting time to extraordinarily detailed or complicated indexing, or to elaborate parsing schemes that refine verbal searches, it might be better to concentrate on indexing the basic elements of an image and rely on scanning, or browsing, to make the fine distinctions."[56] Bates writes, "Most current information systems require that the searcher generate and input everything wanted. People could manage more powerful searches quickly if an initial submitted term or topic yielded a screen full of term possibilities, related subjects, or classifications for them to see and chose from."[57] This approach and the hybrid verbal-visual approach are increasingly appearing in the literature and may address the problem of searchers' inability to specify, verbally, a particular view type.

Evaluation of Search Performance in Information Retrieval Systems

Recall and precision are the conventional measures of retrieval effectiveness. Recall represents that fraction of all relevant items correctly retrieved, whereas precision represents the fraction of relevant items among the items retrieved.[58] Described alternatively, "Failure to collocate all works on a given subject is a recall failure; failure to provide retrieval sets of a reasonable size is either a precision failure or a generality failure. Recall failures occur in cases of synonymy, that is, when subjects are given more than one name; precision failures where there is homonymy; or where the naming is not specific enough."[59] One can infer, then, the important role of thesauri in matching query terms to indexed terms and how this can affect retrieval effectiveness.

Soergel investigated the characteristics of indexing that affect retrieval effectiveness and found that the degree to which searchers can adapt their queries to the retrieval system corresponds closely to achieving satisfactory results.[60] Searchers may seek to increase precision by invoking inclusive searching or using a role indicator to specify a term's association to a given facet or data element. Soergel argues that hierarchically structured indexes that reveal the internal logic of the search system to the user and permit automatic "explosion" of query terms appear to be correlated positively to retrieval.

Heidorn, although referring specifically to indexing colors using the *AAT*, proposes a search system functionality that might be applied more generally to any type of query term. He writes, "should the search fail, the system should be able to automatically relax the matching constraints by moving up to the base term (e.g., pale blue to blue) and perform a search with that high-level term. If that fails, the system might use OR to group all terms having the base term blue."[61]

Kim and Kim propose a knowledge-based information-retrieval system that uses a hierarchical thesaurus and measures the conceptual distance between a query term and a digital object referenced in an online catalog. "Since the IR problem is the identification of those items that contain information pertaining to a user query, the relationships specified in a thesaurus are very valuable in deciding the relevance of an object to a given query."[62] Their model would accommodate both Boolean operators (used by searchers) and term weights (assigned by the system and reflecting a term's position in the hierarchical thesaurus), both of which would contribute to the calculation of conceptual distance, and thus to search system performance.

Classification is an indexing device intended to improve retrieval. Mills describes classification as applied to subject content and "imaginative content," where works of fiction, poetry, drama, and the visual arts represent the latter.[63] Faceted classification offers "highly structured maps of knowledge"[64] to assist searchers and the search system in bridging the gap between the searchers' cognitive perception of their needs and the system's logical interpretation of query terms.

Mills distinguishes between "general" and "special" classifications, the latter intended for a specific field of knowledge. His BC2 represents the former, which will arguably become increasingly important as digital image collections are queried by searchers from diverse disciplines in unconventional ways.

One feature that may greatly improve the searchability of a digital image collection is an alphabetical index that serves as a link between the users' search terms and the system's index terms. *AAT* and *MEDLINE* both offer this feature, which is particularly powerful because it reveals the implicit classification and structure of the search system and offers users broader or narrower terms that they might not have considered initially.

Indexing

Indexing, as used in this article, refers to entries in a catalog intended to be used as access points by the end-user. For digital images of cultural heritage objects, these include both standard descriptive data and subject terms.[65] The former, similar to the data in a bibliographic record, are most often invoked for known-item searches. The purpose of indexing is, broadly speaking, to help searchers identify resources efficiently by organizing resources in a defined order. Mills describes indexing as a function designed for "locating and relating."[66]

Indexers must analyze an image, identify its significant access elements, then translate these into the indexing language, often using controlled terms like subject headings, name authorities, or terms selected from a structured vocabulary like a thesaurus, though sometimes using natural-language terms. Often, the terms of a controlled vocabulary are not the most common terms that users would choose "off the top of their heads." Indexed terms (especially for images) also tend to be more specific; cataloging rules have long advocated the use of specific terms to describe a subject. Catalog users, however, often search using broad terms, especially in the early stage of a search.[67]

Researchers have attempted to identify a core set of image attributes that would in combination provide satisfactory access to digital image collections for a range of user types. Jörgensen suggests four perceptual classes—objects, people, color, and location—which, along with interpretive elements like meaning and narrative content, and "reactive" elements like conjecture and emotions, are likely to represent those indexing elements needed for intellectual access.[68]

Shatford Layne discusses the importance of image attributes as access points and of image groupings, created using these attributes' values, for image searchers. She describes four attribute categories: biographical (an image's title, creator, provenance, date), subject, exemplified, and relationship. These attributes might serve as a guideline at the time of indexing to ensure that all significant data elements have been recorded. Indeed, guidelines that prescribe the indexing of specific facets for every image may offer a better solution. She, along with many other researchers, called for more studies focused on image search behavior and evaluation of usefulness of search results.[69]

Traditionally, expert opinion has tended to place the burden of subject access on the cataloger or indexer. Hourihane notes that the "cataloguer or iconographer acts as a conduit between the work of art (or visual image surrogate of it) and the end-user or researcher."[70] Certainly, an appropriate choice of subject terms will improve access if the user's terminology and conceptualization match exactly those of the indexer. However, this has proven to be a less than common situation, and exact matches are rare. Indeed, Brown reports that 65-80 percent of subject terms fail to match indexed subject headings, and that nearly 50 percent of subject searches return no results at all.[71]

The assumption that thorough subject indexing using authoritative and content-appropriate vocabularies will necessarily lead to effective searching and retrieval for the user must be challenged.[72] Bates reports that many studies have offered empirical (and counterintuitive) evidence demonstrating that searchers and indexers use a wide variety of terms to describe the same document or image with minimal overlap.[73] And yet, indexing schemes are designed on the assumption that consistent application of subject terms will improve access.

Since subject indexing is known to be inconsistent, subjective, and costly, would we not be wiser to compensate for incomplete or erroneous indexing by creating powerful thesauri and shifting the burden to partially automated systems? Rather than trying to anticipate the conceptual links and choices of a seemingly infinite variety of users, we might design systems that process queries through a complex network of hierarchical and syndetic relationships that allow users to input their own natural-language search terms and discover and refine their searches using an intuitive and transparent interface, rather than forcing them to comply with a rigid system and unfamiliar vocabulary.

Bates, in an article on indexing and access for digital collections, describes several factors that designers of access systems should consider: human factors, database factors, and domain factors. With respect to the human factors, Bates remarks that although "it is commonly assumed that indexing and searching are mirror images of each other...this is only superficially a symmetrical relationship." In fact, "The user's experience is phenomenologically different from the indexer's experience" given that the "user's task is to describe something that, by definition, he or she does not know."[74] The indexer, of course, works directly with the document or image and may consult additional sources when assigning terms for access, such as the scope notes in a thesaurus or subject-heading list. While some indexers may try to anticipate the types of queries that searchers will pose (adopting Fidel's "user-oriented" approach or Soergel's "descriptor view"[75]), others may focus exclusively on the material at hand and strive instead to reflect its content accurately (adopting Fidel's "document-oriented" approach or Soergel's "entity view"). The use of thesauri, which enables matching syntactically or semantically different terms, may represent one solution to this cognitive and linguistic gap between indexer and searcher.

Bates also explains how the statistical properties of information systems might inform indexing and design decisions. For example, Zipfian distributions reveal that there are generally a limited number of terms used with high frequency and a larger number of terms used with low frequency; this is true of terms used in both indexing and querying. "These distributions are quite robust and defy efforts of thesaurus designers and indexers to thwart them."[76]

Studies have also revealed that domain-specific vocabularies correlate to intellectual access. Bates reports that Wiberley found the vocabulary of the humanities to be more precise and unique than many had assumed. Other studies, including Bates's study at the Getty Information Institute mentioned above, supported this finding and confirmed that the vocabulary commonly used by searchers in the humanities did not match that used by searchers in the science disciplines; likewise, the paradigms and problem conceptualization of scholars in different disciplines also vary. As a result, Bates suggests that rather than ignoring or trying to overcome the limitations of these statistical patterns and user behaviors, indexers might better focus on the core terms of a given knowledge domain, those terms used most frequently, and spend less time indexing less frequently invoked terms.

Multiple indexing vocabularies and application rules present special problems for cross-collection searching. Turner recommends adopting an indexing format that "allows for both a shared general record and for specific additions made to specialized collections...our energies may be better spent in attempting to establish translation protocols between retrieval systems than in attempting to bend the realities of each collection into a standardized surrogate model."[77]

Suggestions for Additional Research

Most of the researchers cited in this paper call for further investigation of searchers' behavior, in particular its cognitive aspects, and information-retrieval system performance, in particular failure analysis (especially given that most research to date has focused on measures of success and user satisfaction). Additional research addressing how indexers assign subject terms to images, the correlations and gaps between indexed and query terms, and the development of semantic models for analyzing these would contribute to a more cohesive framework for conceiving and implementing systems that provide intellectual access to digital image collections.

Bates and others offer interesting insights into alternate interface designs for online searching using thesauri. Further research might include usability testing of different methods for making a system's internal structure transparent to users. How might a search system automatically prompt users to navigate to and select from among authorized terms? What is the most intuitive way to display results?

It would be useful to investigate further the possibilities of automatic query expansion based on the relationships of indexed terms as defined by semantic networks. Would a search system be able to modify and iterate a query as an expert user does? Bearman notes that "indexers traditionally assign the narrowest applicable term,"[78] yet users may not be searching with such precision. This asymmetry calls for an analysis of the technical feasibility of query expansion to higher hierarchical levels when NTs or natural-language terms return no results. In fact, Greenberg cites studies that present empirical evidence of improved retrieval using query expansion via broader terms, and she suggests that ample research findings exist for more focused research.[79] Additional study of iterative searches and query modifications as performed by different classes of users might also contribute to the literature on effective system design.

More broadly, one might search for the design principles that would maximize, from the searcher's perspective, the advantages of collections indexed using a structured vocabulary or thesaurus. What kinds of searches should be possible? Which ones lend themselves best to structured vocabularies? How can systems most efficiently and accurately convert natural-language queries to indexed terms? Is there a difference, from the searcher's point of view, between systems that make explicit a search result derived from a variant term and those that do not? Standard definitions and applications of thesaural codes would also improve the functionality of thesauri in online searching.

If it is futile to attempt to anticipate user queries, how can we develop thesauri that will optimize the general term-specific term trade-off and work with, rather than ignore, the Zipfian distributions described in previous research? What types of training programs can we develop that will equip indexers with the skills and knowledge needed to minimize inconsistencies and close the gap between their word choices and those of searchers? As the types of users searching digital image collections of cultural heritage materials become more diverse, representing a variety of knowledge domains, how can we create cross-disciplinary search systems that might integrate multiple vocabulary sources to ensure both general and domain-specific terminologies?

Finally, for those working with non-Western cultural heritage materials, multi-lingual thesauri might help address the nagging inconsistencies of transliteration and the use of diacritical marks. Furthermore, currently available thesauri, like *ICONCLASS* and those developed at the Getty, would benefit from the addition of specialized terms used to describe objects from non-Western traditions.

These questions and suggestions for further research reveal the breadth and complexity of this exciting topic. Much important work has been done, laying down the foundation for deeper insights, more powerful tools, and, ultimately, increased user satisfaction. Some day searchers, using a thesaurus as a pass-key, may gain full entry to our digital image collections and enjoy the rich benefits of our shared cultural heritage.

Notes

1. Kodak stopped manufacturing slide projectors in June 2004 and rumors about the potential demise of slide film are circulating; moreover, faculty, students, and the public are increasingly favoring digital image technology over conventional film.

2.This term refers to works of fine art, architecture, and other culturally and historically valuable materials.

3. James M. Turner, "Subject Access to Pictures: Considerations in the Surrogation and Indexing of Visual Documents for Storage and Retrieval," *Visual Resources* 9 (1993): 262.

4. Sara Shatford, "Analyzing the Subject of a Picture: A Theoretical Approach," *Cataloging & Classification Quarterly* 6, no.3 (1986): 60.

5. Corinne Jörgensen, "Attributes of Images in Describing Tasks," *Information Processing and Management* 34, no.2/3 (1998): 162.

6. Hsin-Liang Chen and Edie M. Rasmussen, "Intellectual Access to Images," *Library Trends* 48, no. 2 (1999): 293.

7. Shatford, "Analyzing the Subject of a Picture," 50-52.

8. A.E. Cawkell, "Picture-Queries and Picture Databases," *Journal of Information Science* 19, no.6 (1993): 413.

9. Elaine Svenonius, "Access to Nonbook Materials: The Limits of Subject Indexing for Visual and Aural Languages," *Journal of the American Society for Information Science* 45, no.8 (1994): 600.

10. Svenonius notes that some visual symbolisms might be categorized in a lexicon of forms or iconographic elements, and that these forms and their meanings can be analyzed and interpreted as subjects. Nevertheless, she admits that there can be emotive, expressive elements of a work of art "that cannot be captured by words" and that nonrepresentational works (indeed even poetry) can be utterly and deliberately devoid of the visual and iconographic elements that indexers analyze and convert into subject terms.

11. Svenonius, "Access to Nonbook Materials," 601.

12. Paule Rolland-Thomas, "Thesaural Codes: An Appraisal of Their Use in the Library of Congress Subject Headings," *Cataloging & Classification Quarterly* 16, no.2 (1993): 78.

13. David Bearman, "Thesaurally Mediated Retrieval," Visual Resources 10 (1994): 296.

14. Precoordinated terms present problems for the back-end of a search system; the system must be able to parse a string and to identify its different elements. Many search interfaces for online databases and catalogs circumvent the need for search strings by providing searchers with multiple qualifiers, each matching a specific facet.

15. Qualifiers might be used, for example, to distinguish homographs appearing under different facets.

16. Jane Greenberg, "Intellectual Control of Visual Archives: A Comparison Between the Art and Architecture Thesaurus and the Library of Congress Thesaurus for Graphic Materials," Cataloging & Classification Quarterly 16, no.1 (1993): 94.

17. Precoordination is a method of indexing in which multiple concepts are combined in advance by the indexer to form subject headings (such as *LCSH*) or descriptors assigned to items to facilitate the retrieval of information on complex subjects. Postcoordination is a method of indexing in which the subject headings or descriptors assigned to the items represent simple concepts that the user must combine at the time of searching to retrieve the information. See Joan M. Reitz, *ODLIS: Online Dictionary for Library and Information Science* (Westport, CT: Libraries Unlimited), http://lu.com/odlis/odlis_p.cfm. Postcoordinated terms are more amenable to nonexpert searchers who may not be familiar with the syntax of precoordinated terms.

18. Toni Petersen, "Developing a New Thesaurus for Art and Architecture," *Library Trends* 38, no.4 (1990): 657.

19. Jack Mills, "Faceted Classification and Logical Division in Information Retrieval," *Library Trends* 52, no.3 (2004): 551.

20. Marcia J. Bates, "Indexing and Access for Digital Libraries and the Internet: Human, Database, and Domain Factors," *Journal of the American Society for Information Science* 49, no.13 (1988): 1,201.

21. Cawkell offers a dissenting voice. "In the case of images, the use of thesauri to control inconsistency is not effective due to the individual responses prevalent in human reactions to visual materials…if inconsistency is to be overcome, system designers will need to relinquish the idea of the utility of using words to index non-verbal understanding." (Cawkell, "Picture-Queries and Picture Databases," 413).

22. M.J. Bates, "The Getty End-User Online Searching Project in the Humanities; Report No.6: Overview and Conclusions," *College & Research Libraries* 57 (1996): 520.

23. Mills, "Faceted Classification and Logical Division in Information Retrieval," 541-42.

24. Ibid., 565.

25. Edward T. O'Neill and Lois Mai Chan, "FAST (Faceted Application of Subject Terminology): A Simplified Vocabulary Based on the Library of Congress Subject Headings," *IFLA Journal 29* (2003), 4.

26. Indeed, bias inherent to many available vocabularies (*LCSH, AAT, ICONCLASS*) has motivated the creation of local thesauri for terms describing works of non-Western art and architecture.

27. Rolland-Thomas, "Thesaural Codes," 78-81.

28. This might be an effective way to link terms in the *AAT* that appear under different facets but relate to the same work. For example, the subject term "madrasa" might be listed with related keywords (which are neither BTs, NTs, RTs for "madrasa" but are related, rather, to the concept of "madrasa"), such as Islamic religious education, waqf, four-iwan plan, Islamic law, multi-purpose buildings, or communal living. With simple keyword searching, the multiple subjects associated with a work may be redundant in the hierarchical sense, so combining the power of a carefully structured vocabulary with a list of related keywords might allow searchers to more easily modify their searches by offering both immediate access to BTs, NTs, RTs, and to related keywords that do not fit the strict definition of relatedness stipulated by standard thesaurus design.

29. Rolland-Thomas, "Thesaural Codes," 82.

30. Ibid., 75.

31. Ibid., 87.

32. O'Neill and Chan, *FAST*, 4.

33. Richard Brilliant, "How an Art Historian Connects Art Objects and Information," *Library Trends* 37, no.2 (1988): 120.

34. Helene Roberts, "'Do You Have Any Pictures of…?': Subject Access to Works of Art in Visual Collections and Book Reproductions," *Art Documentation* 7, no.3 (1988): 87.

35. Bates, "Getty End-User," 520-21.

36. Ibid., 519.

37. Sara Shatford Layne, "Some Issues in the Indexing of Images," *Journal of the American Society for Information Science* 45, no.8 (1994): 587.

38. John Sutherland, "Image Collections: Librarians, Users, and Their Needs," *Art Libraries Journal* 7, no.2 (1982): 44-45.

39. P.G.B. Enser, "Pictorial Information Retrieval," *Journal of Documentation* 51, no.2 (1995): 155.

40. Samantha Kelly Hastings, "Query Categories in a Study of Intellectual Access to Digitized Art Images," *Proceedings of the 58th Annual Meeting of the American Society for Information Science* 32 (1995): 7.

41. Colum Hourihane reports that Enser and Armitage found the most used types of search terms to be creator name and subject (Colum Hourihane, "It Begins with the Cataloguer: Subject Access to Images and the Cataloguer's Perspective," in *Introduction to Art Image Access: Issues, Tools, Standards, Strategies*, ed. Murtha Baca [Los Angeles: Getty Research Institute, 2002]), 40. Other researchers have also developed carefully documented theories that name specific image attributes as

preeminent search terms. Yet, given that these studies have produced contradictory results, it would appear unwise to attempt to predict those image attributes most likely to be queried. See also Bates, "Getty End-User"; Turner; Roberts; Enser; Mary E. Brown, "By Any Other Name: Accounting for Failure in the Naming of Subject Categories," *Library and Information Science Research* 17, no.4 (1995): 347-385; and Micheline Hancock, "Subject Searching Behaviour at the Library Catalogue and at the Shelves: Implications for Online Interactive Catalogues," *Journal of Documentation* 43, no.4 (1987): 303-21.

42. Mary Dykstra, "Subject Analysis and Thesauri: A Background," *Art Documentation* 8, no.4 (1989): 173.

43. Shatford, "Analyzing the Subject of a Picture," 43-45. Svenonius and Shatford both assert that indexing at the iconographic level is problematic and vulnerable to inconsistencies.

44. Ibid., 47.

45. Sara Shatford Layne, "Subject Access to Art Images," in *Introduction to Art Image Access: Issues, Tools, Standards, Strategies*, ed. Murtha Baca (Los Angeles: Getty Research Institute, 2002), 12.

46. Brown, "By Any Other Name," 347-48.

47. Marcia J. Bates, "Subject Access in Online Catalogs: A Design Model," *Journal of the American Society for Information Science* 37, no.6 (1986): 368-73.

48. Ibid., 358.

49. Ibid., 368.

50. Ibid., 361.

51. Bearman, "Thesaurally Mediated Retrieval," 296-301.

52. Jane Greenberg, "Automatic Query Expansion via Lexical-Semantic Relationships," *Journal of the American Society for Information Science* 52, no.5 (2001): 402.

53. Shatford Layne, "Subject Access to Art Images," 4.

54. Dykstra, "Subject Analysis and Thesauri," 173.

55. See Turner; Cawkell; and Graeme Baxter and Douglas Anderson, "Image Indexing and Retrieval: Some Problems and Proposed Solutions," *New Library World* 96, no.1123 (1995): 4-14.

56. Shatford Layne, "Some Issues in the Indexing of Images," 586.

57. Bates, "Indexing and Access for Digital Libraries and the Internet," 1,202.

58. Dagobert Soergel, "Indexing and Retrieval Performance: The Logical Evidence," *Journal of the American Society for Information Science* 45, no.8 (1994): 589-99.

59. Svenonius, "Access to Nonbook Materials," 601.

60. Soergel, "Indexing and Retrieval Performance."

61. P. Bryan Heidorn, "Image Retrieval as Linguistic and Nonlinguistic Visual Model Matching," *Library Trends* 48, no.2 (1999): 319-20.

62. Young Whan Kim and Jin H. Kim, "A Model of Knowledge Based Information Retrieval with Hierarchical Concept Graph," *Journal of Documentation* 46, no.2 (1990): 114.

63. Mills, "Faceted Classification and Logical Division in Information Retrieval," 543. Imaginative content consists of both pre-iconographic and iconographic terms.

64. Ibid., 547.

65. Indexing images can also involve the use of visual thesauri that describe visual features in terms of pixels. While useful in some contexts, content-based indexing (as opposed to textual, concept-based indexing) is not refined enough to be effective for many types of image searches. Whether used alone, or in combination with visual information retrieval, language-based indexing, particularly for semantic content, is critical.

66. Mills, "Faceted Classification and Logical Division in Information Retrieval," 542.

67. Shatford Layne, in "Subject Access to Art Images," 2, asserts that an image is unlikely to represent "religious buildings" and rather will depict a particular building or building type.

68. Jörgensen, "Attributes of Images in Describing Tasks," 168, 172.

69. Shatford Layne, "Some Issues in the Indexing of Images," 587.

70. Hourihane, "It Begins with the Cataloguer," 40.

71. Brown, "By Any Other Name," 34.

72. Enser goes so far as to dismiss linguistic access to images altogether. "We are drawn to the conclusion that the attempt to satisfy queries by matching their linguistic form against the linguistic identifiers, in the form of indexing terms, titles, and captions, attached to images within a collection offers little promise as an effective pictorial information retrieval procedure." (Enser, "Pictorial Information Retrieval," 156).

73. Bates, "Indexing and Access for Digital Libraries and the Internet," 1188-90.

74. Ibid., 1,186.

75. Raya Fidel, "User-Centered Indexing," *Journal of the American Society for Information Science* 45, no.8 (1994): 572-76; Soergel, "Indexing and Retrieval Performance." Fidel describes two approaches to indexing, document-oriented, and user-oriented and notes that the latter cannot be developed well until further research elucidates user behavior. Fidel asserts that a document-oriented approach typically focuses on the processes of subject and concept analysis, description, and selection of appropriate indexing terms, often in accordance with pre-defined rules or guidelines that prescribe syntax, preferred terms, level of detail, and coordination or combination of approved terms. This approach favors the accurate representation of a document's content. An alternate approach would favor accurate matching of indexed terms to query terms, placing the primary emphasis on users of information retrieval systems, their needs, and natural-language information requests. In this case, indexing terms are derived from likely or known user queries, in effect, working backwards from the search term to the indexing term.

76. Bates, "Indexing and Access for Digital Libraries and the Internet," 1,194.

77. Turner, "Subject Access to Pictures," 254.

78. Bearman, "Thesaurally Mediated Retrieval," 296.

79. Greenberg, "Automatic Query Expansion via Lexical-Semantic Relationships," 410.

Johanna Woll, Islamic Image Collection Specialist, Massachusetts Institute of Technology, Cambridge, jwoll@mit.edu