5

An analysis of search terminology used by humanities scholars: The Getty Online Searching Project report no. 1

ABSTRACT

The Getty Art History Information Program carried out a two-year project to study how humanities scholars operate as end users of online databases. Visiting Scholars at the Getty Center for the History of Art and the Humanities in Santa Monica, California, were offered the opportunity to do unlimited subsidized searching of DIALOG® databases. This first report from the project analyzes the vocabulary terms twenty-two scholars used in their natural language descriptions of their information needs and in their online searches. The data were extracted from 165 natural language statements and 1,068 search terms. Vocabulary categories used by humanities scholars were found to differ markedly from those used in the sciences, a fact that imposes distinctive demands on thesaurus development and the design of online information systems. Humanities scholars searched for far more named individuals, geographical terms, chronological terms, and discipline terms than was the case in a comparative science sample.

First published as Bates,[1] M. J., Wilde,[2] D. N., & Siegfried,[2] S. (1993). An analysis of search terminology used by humanities scholars: The Getty Online Searching Project report no. 1. *The Library Quarterly, 63*(1), 1–39.

[1] Graduate School of Library and Information Science, University of California, Los Angeles.

[2] The Getty Art History Information Program.

The analysis provides substantial support for the growing perception that information needs of humanities scholars are distinct from those of scholars in other fields, and that the design of information-providing systems for these scholars must take their unique qualities into account.

Introduction

Stimulated by scholars' increased access to online databases such as those offered by Dialog Information Services, the Getty Art History Information Program launched a two-year program in 1989 to study how humanities scholars, engaging in their characteristic modes of research, use online databases when they are given the chance to do unlimited searching, unconstrained by cost. The study was intended to probe a number of aspects of the scholars' experiences with online searching, including their reactions to the use of the online databases, the role the searching had in their research work, their search techniques and learning curve, their queries, and the search terms they used. The results of the study are to be reported in a series of articles, of which this is the first.[3]

The data analyzed for this article are the natural language statements the scholars gave describing or commenting on their queries, and the search terms used by the scholars in their online searches. The analyses reported here were prompted by the observation that the wording the scholars used in their natural language statements contradicted some prevailing assumptions in information science about subject terminology in online searching. This article identifies, quantifies, and discusses the principal categories of search terms the scholars used in both their natural language statements and their online search statements to DIALOG.

Background

Empirical research into information-seeking behavior among members of various academic and research communities focused almost exclusively on engineering and the sciences during the 1960s, and on the social sciences in

[3] Marilyn Schmitt conceived the Getty Online Searching Project; she designed the study together with Susan Siegfried and Deborah N. Wilde. Siegfried and Wilde carried out the collaborative project plan and oversaw the gathering of data throughout the project. Marcia Bates analyzed the data, formulated conclusions, and contributed insights from the discipline of information science. Research assistant Vanessa Birdsey coded the data into categories developed by Bates. Jeanette Clough and Katherine Smith transcribed the DIALOG search statements for subsequent analysis.

the 1970s. Although Sue Stone reviews a number of studies done before 1980, she points to the "relative neglect" of the humanities until the late 1970s. Her review (Stone, 1982) and the work begun in the late 1970s by the Centre for Research on User Studies (for example, Corkill & Mann, 1978) mark the beginning of a modern era of interest in humanities scholars' information seeking on the part of researchers in library and information science.

During these decades, reviews, research projects, and commentaries drew attention to the unique characteristics of the information-seeking behavior of humanities scholars (Broadbent, 1986; Case, 1991; Corkill & Mann, 1978; Gould, 1988; Guest, 1987; Lowry & Stuveras, 1987; Rahtz, 1987; Schmitt, 1988, 1990; Stam, 1984; Stam & Giral, 1988; Stielow & Tibbo, 1988; Stone, 1982; Tibbo, 1989; Wiberley, 1991; Wiberley & Jones, 1989). While many of these studies dealt with the research behavior of scholars in general, few paid attention to these scholars' use of online databases. However, during the last few years this lack has begun to be remedied. Serious attention is now being given to online searching in the humanities: its particular requirements, problems with databases and their vendors, and specific search techniques (Boyles, 1987, 1988; Crawford, 1986; Everett & Pilachowski, 1986; Katzen, 1986; Lehmann & Renfro, 1991; Loughridge, 1989; Mackesy, 1982; Ross, 1987; Ruiz & Meyer, 1990; Stern, 1988; Walker, 1988, 1990; Walker & Atkinson, 1991).

A few empirical studies have examined actual uses of online databases by or for scholars in the humanities. Janice Woo (1998) studied how three graduate students at Columbia University made online use of the *Avery Index to Architectural Periodicals*. Sylvia Krausse and John Etchingham elicited the reactions of scholars to database searching when a grant subsidized the cost of their searches (Krausse & Etchingham, 1986). Jan Horner and David Thirlwall tested several hypotheses regarding uses of online searching by social science and humanities scholars (Horner & Thirlwall, 1988), and Jitka Hurych (1986) analyzed formal online search requests to compare the use of online search services across disciplines (sciences, social sciences, and humanities).

Of particular relevance to this article are studies of humanities vocabulary and indexing. Stephen Wiberley, who has conducted two large research projects on indexing vocabulary in the humanities, studied the vocabulary of encyclopedias and dictionaries (Wiberley, 1983) and of abstracting and indexing services (Wiberley, 1988) in the humanities to test the truth of the cliché that the vocabulary of the humanities is vague and imprecise. He found, on the contrary, that much of the vocabulary in the humanities, which consists largely of names of individuals and works, is in fact very precise.

Geraldene Walker studied how certain trial subject terms were distributed across nine databases available on DIALOG (Walker, 1990; Walker & Atkinson, 1991). Bella Weinberg (1988) argues that the scholar is ill served by indexing systems that deal only with the topic, or "aboutness," of materials. She states that scholars need, in addition to topic, "comment" information that describes point of view and/or the specific argument or theory presented. Some of this literature will be examined in more detail later.

Project methodology

Population

All Visiting Scholars during 1988–89 and 1989–90 at the Getty Center for the History of Art and the Humanities in Santa Monica, California, were invited to participate in the project, but not all chose to do so.[4] Eleven of the fifteen scholars visiting the center during 1988–89 agreed to participate in the project, and two spouses (scholars in their own right) also participated. Twelve of the eighteen scholars during 1989–90 agreed to participate, along with three spouses. Though the numbers of participants for the two years add to twenty-eight, one individual in the 1989 group stayed for the second year, for a total of twenty-seven different people.

For purposes of this study, "participation" was defined as taking the DIALOG training. Some scholars did no searching, however, or searched only with help from an experienced searcher. We were interested only in those individuals who searched on their own at some time during their stay at the Getty: that is, those who produced natural language statements and search statements while acting as true end users. Eleven participants conducted such unassisted searches in 1989, and twelve did so in 1990. Since one person stayed the second year, twenty-two different people produced the data analyzed here. From now on, discussion of the scholars producing data for this study will refer to this smaller group. One of this group did just one unassisted search but made no natural language statement, for a final total of twenty-one individuals producing natural language statements and twenty-two producing online search statements.

In the second year of the experiment, one individual spent more than five times as many hours searching as any of the other scholars—in

[4] The J. Paul Getty Trust is a private operating foundation. Two of its programs are the Getty Art History Information Program and the Getty Center for the History of Art and the Humanities. These two entities collaborated on this project.

fact, more time than all the others put together. Since the vast majority of participants did relatively little searching, the various analyses in the study were based on the early hours of the scholars' experience as searchers. We decided to end the analysis of the prolific scholar's searching at a point beyond the amount done by everyone else but well before the end of his total searching time—on the grounds that analyzing later stages of searching with a sample of one would be of little use. Thus, total numbers of natural language statements and search terms for this individual in the analyses in this article are based on the subset of material actually analyzed, not the entire record.

The thirteen male and nine female scholars came from France, Germany, Great Britain, Hungary, Italy, and the United States. Eight were native English speakers; the nonnative speakers' command of English ranged from adequate to excellent. The scholars' research interests included the history of art and architecture, film history, social history, philosophy, comparative literature, classics, the history of music, and social and cultural anthropology. The group comprised university professors, independent scholars, a curator, an architect, postdoctoral scholars, and doctoral candidates.

Training, setting, and other arrangements

Participants were given one day's training by Amy Greenwood, a DIALOG staff trainer, in late January and early February for the 1989 group, and in November of the second project year for the 1990 group.[5] The scholars then had twenty-four-hour-a-day access to a workstation in the Getty Center Library near their offices until they left in the summer. Next to the workstation were placed documentation for DIALOG, as well as thesauri and word lists for several arts and humanities databases. The latter included *RILA (International Repertory of the Literature of Art) Subject Headings, Architectural Keywords,* and *Historical Abstracts Index.*

In preparation for this project, a program was written to capture a complete transaction log of the DIALOG searches done by the scholars at the workstation. These data were captured for the study with the permission of both Dialog Information Services and the study participants. Scholars were instructed to print out all desired search results at the terminal, rather than having them sent from DIALOG. Consequently, we have a complete record of their entire searches—both search statements and results.

[5] One scholar arrived at the Getty Center after the training had ended and so was trained individually by Jeanette Clough of the Getty Center, rather than with the group.

The Art History Information Program arranged for a limited DIALOG account, which gave the scholars access to a large subset of about sixty of the DIALOG databases. Databases in the package, drawn from the social sciences, arts, and humanities, included all those thought to be of interest to arts and humanities scholars, such as *Art Literature International* (*RILA*), *The Architecture Database* (*RIBA*), and *Historical Abstracts*. (Although *RILA* has since been superseded by the bilingual *Bibliography of the History of Art* [*BHA*], the database is still accessed through DIALOG as *Art Literature International*.) Bibliographic databases covered journal articles, books, and dissertations. Some directories, for example, *Marquis Who's Who*, and some full-text databases, for example, *Academic American Encyclopedia* (no longer available on DIALOG), were also included. DIALINDEX®, the database of DIALOG databases, was included, as were citation databases, for example, *Arts & Humanities Search®*.

Participants were encouraged to make an appointment for an "assisted search" at some time during the months after the training. In other words, an experienced online searcher would answer questions and help in any way desired while the scholar searched. During the first year, six scholars requested an assisted search during the spring of 1989, and one of these had a second one as well. In 1990 one scholar had three assisted searches, three had two, five had one, and six had none.

Some scholars were also offered the opportunity to have an experienced searcher do a "comparative search." Essentially, the expert searcher redid one of the searches the scholar had already done. The scholar first submitted a written search request to the expert, who then conducted the search (without discussing with the scholar what he or she had done). The results of the comparative searches were discussed in an interview with the scholar. In 1989 seven scholars requested comparative searches (performed by Kathleen Salomon of the Getty Center Library). These comparative searches took place after the scholars had done most of their searching for the year; none occurred earlier than May 4, 1989. They were discontinued in the second year.

Two group review sessions were offered during the second year. Three people attended the first one in January 1990, and five people (including one who also attended the first session) attended the second one in March. (In both years help with assisted searches was provided by Jeanette Clough of the Getty Center Library, who also conducted the two group review sessions.) Overall, the project was designed to encourage scholars to do their own searching; generally, assistance was made available only through the DIALOG help line, not locally.

Analysis of natural language statements

Methodology

Introduction As a part of their participation in the study, the scholars were asked to type in a description of their query at the beginning of each search they performed. As these descriptions were not preceded by a DIALOG command, they were meaningless to DIALOG and had no effect on the online search itself. However, this text was captured by the program that recorded a complete transaction log of the search. Thus both natural language statements and the entire online search were recorded at the moment of search and made available for later analysis. The natural language statements analyzed in this study consisted of all the descriptions of queries and other comments that the scholars input during their unassisted searches. No natural language statements associated with assisted searches were analyzed; the analysis was restricted to comments searchers made as they worked on their own.

Because the scholars were searching on their own without experimenter prompting, their natural language statements did not always correspond perfectly to their online searches. On occasion scholars entered searches for which they failed to provide a natural language statement. In other cases, they made several comments on a single search or searched the same or a similar query in another search session. Sometimes these repetitions seem to represent a shift from one way of thinking about a search to another approach, and sometimes they entailed the incorporation of new terms.

Categories Several sets of terminological categories were developed and experimented with in order to find those that best revealed the special characteristics of the vocabulary used by humanities researchers in online information seeking. The intent was to discover what was unique about humanities online terminology.

The following categories were selected. They consist of three broad classes divided into subcategories:

1. Type of search need
    a. A specific work or publication
    b. Works or publications by an author—specific item not stated
    c. Works on a subject—all senses of subject, including material about a work or an author
2. Bibliographic features
    a. Bibliographic form of desired materials
    b. Publication date or date range of desired materials

3. Types of subjects
    a. Works or publications as subject
    b. Individuals—all sorts of people, including authors, as well as fictional, mythical, or religious characters
    c. Geographical name
        i. Noun form
        ii. Adjective form
    d. Date or period
        i. Date or date range
        ii. Period
        iii. Time modifier
    e. Discipline
    f. Other proper term
    g. Other common term

Comments on the categories "Type of search need" included three broad types of search likely to be conducted: for a specific work, for any works by an author, and for works on a subject, which may include searches on works and authors *as subjects*. The phrase "work or publication" is used because a searcher may seek either a specific publication or a work of an author that may appear in any number of different editions, translations, and so forth.

An additional category concerned certain bibliographical features—specifically, bibliographic form and publication date or range of publication dates. Bibliographic form means the common forms of publication, such as books, articles, dissertations, and so forth. Publication date refers to the date of any bibliographic entities to be retrieved, not the time period covered in the text of the items.

Finally, the largest set of categories contained types of subject searches. The category of works as subjects may include works of any of the kinds studied in the humanities—literary, artistic, musical, and so forth. Individuals include any real, fictional, or mythical characters. Geographical names include political or historical entities with geographical boundaries, for example, "United States" or "Weimar Republic." "Adjective form" refers to adjectives such as "Dutch."

Date may refer either to specific dates or to date ranges, for example, 1812 or eighteenth century. Period refers to verbal labels for historical periods, such as "Renaissance." While the actual time spans that such terms encompass may be a matter of debate, the terms were included because they are widely and productively used. Time modifier refers to terms such as "early" in "early nineteenth century," which further specify the stated date range or period.

Discipline refers to broad areas of study, such as history, music, or the humanities. The relatively frequent appearance of terms of this sort was one of the surprising results of this study, and a detailed analysis of this category of terms is included below. Since many subjects can be considered "areas of study," this category was coded conservatively: only very broad terms were counted, such as "humanities," down to and including university department–sized areas, such as "art history." "Film," for example, as an area of study, is large enough to merit an entire department on some campuses, but not on others, and so was not categorized as a discipline term.

"Other proper terms" included all proper subject terms not included in any of the above categories of subject, and "other common terms" included all common subject terms not included above. A term was considered proper if it was normally written with uppercase initial letters and common if normally written with lowercase initial letters. Some examples of other proper terms encountered in the study are the following: the "Annunciation," the "Hanging Gardens," and the French title of an individual being researched: "Surintendant des Batiments." Examples of other common terms are "synaesthesia" and "intuition."

A final note is in order on the distinction between the first and third group of categories listed above: type of search need and type of subjects. There is an important distinction between a search for a work *itself* and a search for a work *as subject*. It is one thing to search for Brecht's *Threepenny Opera* itself and quite another to search for books or articles discussing it. Similarly, there is a great difference between searching for works *by* Brecht (author search) and works *about* him (subject search).

This distinction between the work itself and materials about the work has an implication for category definition that might not be immediately evident. In the humanities, scholars may be interested in works of all kinds—paintings, dance and musical compositions, literary works, and so forth. Works of all these types were therefore categorized under works as subjects. On the other hand, when a scholar searches for the work itself in a bibliographic database, that work must be bibliographic in nature: that is, it must be some sort of recorded work of the kind typically contained in library catalogs and databases. Such databases might contain a record for a novel, but not for a painting. So the types of works actually included within the categories of work as type of search need and work as type of subject were necessarily different.

Identification and coding of terms Each of the three broad categories listed above—type of search need, bibliographic features, and type of subject—was coded separately for each natural language statement (NLS). The third category, type of subject, was coded for an NLS only if subject had been identified as a type of search need for that NLS.

One of the complications of an analysis such as this is the question of what constitutes a subject term. Is it a single concept expressed in one or more words, or is it strictly a single word? In library and information science the word "term" normally has the former meaning. If we accept that interpretation, then how do we decide whether a phrase contains one or more concepts? With these data, it was very difficult to isolate what the user considered a unitary concept. As noted earlier, searchers were given great flexibility in making comments and describing searches. They wrote in ordinary narrative style, sometimes quite colloquially, sometimes cryptically. And such difficulties are certainly not limited to this study. Even when experts identify concepts as a part of thesaurus development, rules for different thesauri and term lists can vary in this regard. Furthermore, even the application of a given set of rules, such as those in a thesaurus standard, can be complex and difficult. As one standard notes, "The establishment of procedures for dealing consistently with compound terms introduces one of the most difficult areas in the field of subject indexing" (International Organization for Standardization, 1986, p. 9). Finally, in online searching, various searchers—or even one searcher under different circumstances—may treat single words and multiword phrases very differently.

This problem is well illustrated by phrases that appeared in three successive natural language statements: "funerary masks," "funerary representation," "representation of Christ." The searcher might or might not have thought of each of these three as unitary concepts. Thesauri also might handle them variously. Would "Christ representations" be a distinct term, like "funerary mask," or would "Christ" and "representation" be treated as distinct concepts?

Similarly, when recording these natural language statements, the scholars were about to do an online search. Would the wise searcher treat each of these three phrases (1) as a descriptor, that is, as a unitary concept, (2) as composed of distinct words that must be combined with Boolean logic, or (3) as a natural language phrase, which must be expressed with proximity operators? The last case represents an in-between situation, in which the searcher finds the phrase meaningful as a phrase, and thus wants the words to be in proximity, but may not expect to find the phrase as a descriptor—that is, recognized as a unitary concept by the database's thesaurus.

This is not to suggest that the scholars, who were inexperienced in online searching, would have considered any or all of these issues but, rather, that even if they had, the matter would remain difficult to interpret. Thus, counting the individual terms in the thesaurus sense in these natural language statements appeared to be problematic, to say the least. More important, the attempt to impose our interpretation on the scholar's terminology could have biased the analysis and reduced the validity of the results.

Thus, we took a different approach. Our fundamental unit of analysis in the natural language statements became the "appearance" of a category in the statement. That is, if some language representing a category appeared anywhere in a natural language statement, whether in one or several words, or in one or several instances, that event, called an "appearance," would be counted as one in the tally. For example, we did not attempt to decide whether "funerary representation" was one or two concepts. Both of these words, "funerary" and "representation," fit the other common category. Therefore, this statement was coded as having one or more instances of the other common category, that is, as one appearance of that category. In like manner, "Greek and Roman libraries" would be counted as one appearance of the geographical (adjective form) category and one appearance of the other common category.

A terminological problem still remains. If "some language" regarding a geographical location appears in a scholar's natural language statement, and we call that event an "appearance" in the tally, we still need to be able to refer to that "some language" in some compact way. Hence in the following discussion of the analysis of the natural language statements, we will use the word "term" with the special understanding that it refers to the "some language" in the definition of "appearance." For example, when we say that a geographical term appeared in an NLS, we mean that some language, in one or more words and in one or more instances, which had some geographical meaning, appeared in the NLS. "Term" and "appearance" are used in these defined senses in tables 2, 3, 4, and 9, and in the accompanying text. "Search term" will have a distinct definition in the later discussion of the analysis of terms used by the scholars in their online search formulations.

Finally, when a specific title of a work was named either as the item being sought or as the subject of interest, the title of the work was coded as a title only; the words in the title were not coded into categories.

Details of wording Scholars were encouraged both to describe their search topics and to make any comments they wished on the search itself. It was felt that anything the searchers chose to say about the search while it was going on might help us understand their use of the online searching capability.

Consequently, wording of the comments is varied in completeness and detail. In each case, the terms of the natural language statements were coded as fully as possible given the information available. Elements of the statements that were comments of one sort or another, not descriptions of current or projected searches, were not coded. A small percentage of the statements contain no codable terms of any kind, as in "I'm continuing," and "I need some info to complete a book review." The second statement failed to make clear what kind of information was needed, so this item was not coded.

In other cases, the information given makes coding possible in some categories, but not others. For example, "Now a broader subject search." Here, the type of search can be coded, but no specific subject terms.

In some cases, a natural language statement is partly a query statement and partly a comment. In these cases the former was coded and the latter ignored, as in the following example: "I will search in the philosopher's index any articles or book written about the problem and the history of imagination, as discussed in the late eighteenth century. I just discovered a book which I did not know of, and now I would like to get to know at least other secondary literature in order to trace further 18th-century material in the libraries." The second sentence was treated as a comment.

Relation of natural language statements to information needs Two final, subtle, but important issues of methodology arise in the coding of the natural language statements. The first has to do with the relationship of information need to online search formulation. Many years ago Robert Taylor (1968) noted that users of information services frequently compromise their real needs when they come to use information services or resources: that is, they formulate their queries according to what they think the system can offer, so the query as presented may differ significantly from the real need.

More generally, in all online searching the information need as it arises in the researcher's mind must be transformed into the query, which is then formulated in search statements understandable to the system. After their day of DIALOG training, the scholars were aware that an information need must be converted into search statements using a rigid syntax and various other rules that differ greatly from common discourse.

Both meaning and syntax are thus liable to change during this process of transformation from need to query to search formulation. What point along this continuum do the scholars' natural language statements represent? Are they closer to the conversational mode the researcher might use with a friend or fellow scholar, or to the rigid syntax required by the online system?

The answer is that the statements appear to be closer to the former than to the latter. The scholars' tone is generally conversational and natural, if somewhat abbreviated. The searchers are probably talking more to those of us conducting the study than to DIALOG, although their wording may already contain some of the compromises needed to communicate with DIALOG. It seems reasonable to assume that the scholars' queries as stated are fairly close to their true needs. Thus in studying the natural language statements we learn of the terminology (and, more generally, the categories of terminology) that humanities scholars use when they discuss the information needs they bring to an online system.

The second issue concerns another aspect of the conversational context. If in their natural language statements the searchers are speaking to those of us conducting the study, they may provide information that they would not normally feel a need to provide in searching queries that were not being observed. In particular, one feature that appeared in some of the statements aroused such a suspicion. Searchers sometimes provided explanatory tags for the personal names they listed in their descriptions of information needs, as in the following examples (tag lines are italicized): "I am looking for articles and books recently written about Victor Cousin, Samuel Taylor Coleridge, and Hegel, *three philosophers of the early 19th century*" and "Henry van de Velde, *Belgian architect and designer*."

Nine percent of the natural language statements studied contained a personal tag line of this sort. Are these tag lines elements that scholars would normally have in mind as their query, or are they added to their usual expression of information need as explanatory elements for the researchers? Whatever the answer to this question, another possibility must be considered as well. Perhaps the searchers are using these tag lines in anticipation of needing such information in their search formulations. In that case, the statements with personal tags would be somewhat closer to search formulations than is the case with other natural language statements.

The current structure of most databases makes it impossible to use the information in the tag lines effectively in a search formulation, at least not in this form. Often the name alone is not only sufficient but in fact is the best way to search on the person of interest. However, one database of central importance in art, *Art Literature International* (previously *RILA*, now superseded by *Bibliography of the History of Art*), uses lengthy tag lines in the descriptors for individuals, for example, "Tissot, James Joseph Jacques, French painter, 1836–1902." So searchers may have added these tag lines in their statements as a result of exposure to this pattern in studying the *RILA* database, either in training or as they began to search.

TABLE 1. *Types of search*

TYPE OF SEARCH	FREQUENCY	%
Work or publication itself	10	6
Materials by an author (specific items not designated)	5	3
Material on a subject	147	89
Both work or publication and subject	1	1
Both author and subject	2	1
TOTAL	**165**	**100**

It is difficult to guess after the fact the source(s) of this tendency and whether these tag lines represent integral elements of the query or supplemental conversational information for the benefit of those conducting the study. Because of this uncertainty, the data on natural language statements were analyzed both with and without the tag lines to see whether similar or different patterns in category frequencies emerged.

Results

Statements Searchers made a total of 188 natural language statements. Subtracting from this number those that did not contain even partial descriptions of information needs or planned searches, we were left with 165 natural language statements, which form the basis of the analysis below.

Distribution of natural language statements by individual The distribution of the number of NLSs across searchers varies widely, with some being quite prolific and others saying, and searching, little. The number of statements per individual ranged from one to thirty-nine, with a median of four. The five most prolific searchers, all with thirteen or more statements, jointly account for 61 percent of all natural language statements. The searcher who carried over from 1989 to 1990 contributed two statements during 1989 and thirteen in 1990. The scholar with thirty-nine statements was the most prolific overall in amount of searching and amount of time spent searching.

Type of search Listed in table 1 are figures for types of search. Since 150, or 91 percent, of the natural language statements indicated a subject search of some kind, it appears that the scholars used these databases primarily

TABLE 2. *Frequencies of subject categories in natural language statements*

SUBJECT CATEGORY	FREQUENCY	PERCENTAGE OF ALL NLSs[a]	PERCENTAGE OF SUBJECT NLSs[b]
Works or publications as subject	8	5	5
Individual as subject	74	45	49
GEOGRAPHICAL NAME:			
Noun form	16	10	11
Adjective form	22	13	15
TOTAL GEOGRAPHICAL TERMS	37	22	25
CHRONOLOGICAL TERMS:			
Date or date range	18	11	12
Period	9	5	6
Time modifier	8	5	5
TOTAL CHRONOLOGICAL TERMS	26	16	17
Discipline	35	21	23
Other proper	11	7	7
Other common	85	52	57

NOTE: Percentages may add to more than 100 percent because an NLS may contain more than one type of term. Since it may also contain more than one type of geographical or chronological term, individual figures within those categories need not sum to the "total" figures for them either.

[a] *Total of 165.* [b] *Total of 150.*

as subject search resources, not for bibliographic verification or as a means of seeking out specific items. They also had access to UCLA's ORION online catalog, and interviews show clearly that they understood that the catalog would work better as a finding device for known citations than would online databases.

Subject categories All figures in table 2 are for appearances in natural language statements of the designated subject category. The first figure is the number of statements containing the designated category; the second figure is that number converted to a percentage of all natural language statements; and the third figure is the percentage of subject search NLSs that the raw number represents. Since natural language statements frequently contain many different kinds of terms, figures can add to far more than 100 percent—but note that all figures are for percentages or numbers of NLSs containing the category, not total mentions of that category. Frequencies for works and individuals are for these *as subjects*, not for the work itself or for the works of an author.

More than one type of date term can appear in the same natural language statement, so the percentage of total NLSs containing date terms is smaller than the sum of the percentages for the subcategories within date, as is also true for subcategories within geographical terms.

Summarizing from table 2, of all subject natural language statements made by these scholars, about half mention one or more individuals or characters, a quarter mention some geographical entity, a sixth mention some date or period, and a quarter mention an intellectual or academic discipline. At the same time, nearly three out of five mention other common subject terms.

These figures begin to highlight some of the differences between the vocabularies of the humanities and of other disciplines. Subject terms falling in our "other common" category predominate in many fields in science and engineering, and in the thesauri for those fields as well. Yet if we turn our figures around, we can say that in the Getty study fully 43 percent of all subject NLSs (100 percent minus 57 percent) make no mention whatever of the kind of subject terms that are the very heart of science and engineering search queries. On the other hand, the humanities queries contain significant percentages of subject categories that are seldom seen in the sciences.

Contrast with science queries To substantiate the point that science queries differ from those in the humanities, we compare the statistics for these natural language statements with those for the search queries used in a major research project performed for the National Science Foundation (NSF) by Tefko Saracevic and Paul Kantor, who reported the results of their research in a series of 1988 articles in the *Journal of the American Society for Information Science*.

One of Saracevic and Kantor's articles includes an appendix reporting the questions used in their NSF study (Saracevic & Kantor, 1988, pp. 195–96). Elsewhere, the authors described these statements as containing "a summary of the text for each question" used in the study (Saracevic & Kantor, 1988, p. 179). The questions were from forty different users, who posed one written question each. The texts of the questions reported are in standard English; they have not yet been converted into search statements and so are roughly equivalent to the statements by users in the present study: that is, they represent a midpoint between original information need and search formulation presented to the system. The NSF statements are more consistent, however, in that all have been converted into a standard format, as in the following examples: "the relationship and communication processes between middle aged children and their parents" (Saracevic & Kantor, 1988, p. 195) and "occurrences, causes, treatment, and prevention of retrolental fibroplasia" (Saracevic & Kantor, 1988,a p. 195).

The forty queries in the NSF study came from many subject fields. (All of the NSF questions would be classified as subject queries according to the definition used in this study.) For comparison with the results in this study, we grouped the NSF queries into three broad classes: (1) arts and humanities; (2) social and behavioral sciences, including management, social work, and other applied social science fields; and (3) natural sciences, engineering, and other applied science fields, including medicine. The number of NSF queries falling into each of the three broad subject areas were two, twenty-two, and sixteen, respectively.

Table 3 displays the frequencies of the major subject categories in this study and in the social and natural science queries of Saracevic and Kantor's NSF study. The two arts and humanities queries were excluded from the latter. The method of counting appearances of term categories that was used in this study was also used in calculating figures for the NSF study.

Though the frequencies in the NSF study are based on fewer statements, the contrast in distribution of subject categories between the studies is so dramatic that we can state with confidence that, as represented in these data, the arts and humanities differ from the sciences in fundamental ways. Other common terms were used in all sixteen natural science queries; only once was a category besides "other common" used. By contrast, fewer than three out of five Getty statements mentioned such other common terms. None of the NSF queries mentioned individuals or characters as subjects, while half of the Getty statements did. None of the NSF queries mentioned discipline, while a quarter of the Getty statements did. Many other contrasts could be drawn.

In fact, after seeing the results in table 3, we returned to the data and made another tabulation. Combining all other types of terms besides other common, that is, all non–other-common terms, only 18 percent of the Saracevic social and natural science queries used one or more non–other-common terms, while 84 percent of the Getty natural language statements did.

So other common terms are only moderately important for humanities queries, while they are the essence of science queries. On the other hand, proper names of individuals and other proper terms, works as subjects, and geographical, chronological, and discipline terms are vital to humanities queries, but much less important in the sciences.

Table 3. Frequencies of subject categories in NSF and Getty studies

	NSF SOCIAL SCIENCE		NSF NATURAL SCIENCE		NSF TOTAL		GETTY TOTAL	
CATEGORY	*N*	%	*N*	%	*N*	%	*N*	%
Total subject queries	22	58	16	42	38	100	150	100
Works or publications as subject	1	5	0	0	1	3	8	5
Individuals as subject	0	0	0	0	0	0	74	49
Geographical name	3	14	0	0	3	8	37	25
Chronological term	1	5	0	0	1	3	26	17
Discipline term	0	0	0	0	0	0	35	23
Other proper term	3	14	1	6	4	11	11	7
Other common term	22	100	16	100	38	100	85	57

NOTE: *Percentages are the percentage of total query statements in each sample in which one or more terms of a given category appeared.*

These extremely marked contrasts suggest that humanities online information seeking differs in fundamental ways from that of the subject domains that research into online searching has hitherto studied in greatest detail. If we are inclined to suspect that the differences in results might be due to some difference in design of the two studies rather than the subject matter, we are reassured by contemplating the two NSF queries in the humanities. Although two queries are too few for proper comparison, it is interesting to note that the pattern exhibited by the two NSF humanities queries is very similar to that of the statements in this study.

While both of the NSF humanities queries mention other common terms, one also mentions geographical and chronological terms, and the other mentions geographical, chronological, discipline, and other proper terms as well. So just two humanities queries contain six appearances of categories other than other common, while all the remaining thirty-eight NSF queries put together contain a total of only nine such instances. Clearly, information queries of the sciences and the humanities differ in fundamental ways.

In fact, we can identify a standard type of arts and humanities query, as exemplified by one each from the NSF and Getty studies: NSF, "meaning of the cat in Italian renaissance (1450–1600) religious paintings" (Saracevic

& Kantor, 1988, p. 196); Getty, “image of the tree in literature, art, science, of medieval and renaissance Europe.”

Further analysis of the Getty/NSF contrast Let us look more closely at the contrast between the Getty and NSF studies to see whether we can discern causes or alternative explanations. One immediate possibility is that the difference may be due chiefly to humanities scholars’ strong interest in research on individuals. Because half of the subject natural language statements in the Getty study concerned individuals and nearly that many contain no other common terms, perhaps these two facts are related and account for most of the differences.

We therefore extracted all the statements that contained no other common terms and analyzed them separately as a group. Of the total group of 165 Getty natural language statements, eighty (or 48 percent) contained no other common terms whatsoever. Seventeen percent of the eighty NLSs without other common terms represented searches for works or authors rather than subjects. Another 64 percent of the eighty were searches for individuals as subjects, and another 10 percent were searches for works or publications as subjects.

So, for a grand total of 91 percent of the cases in which no other common terms were present, the scholar was searching for particular works, for works by an author, or for works or individuals as subjects. We can see from this fact that the large number of natural language statements with no other common terms correlates strongly with the tendency of humanities scholars to study particular works and individuals.

Keep in mind, however, that although this set of statements excludes other common terms, the scholars mentioned types of subject terms other than works and individuals in these statements as well. Five percent used geographical terms, 4 percent chronological terms, 11 percent discipline terms, and 5 percent other proper terms.

On the other hand, often when other common terms do appear, many of the other types of terms that are distinctive to humanities statements appear as well. For example, in twenty-three natural language statements (the seventy-four NLSs on individuals as subject in the entire sample minus the fifty-one on individuals in the “no other common terms” sample), other common terms appear in conjunction with searches for individuals. Similarly, the great majority of geographical, chronological, discipline, and other proper terms also appear in statements in which other common terms appear. In other words, even in those humanities subject searches that more closely resemble science queries because they contain other

TABLE 4. *Frequencies of natural language statement categories with and without personal tags*

	ALL NLSs[a]		NLSs WITHOUT TAGS[b]	
CATEGORY	*N*	%	*N*	%
Type of search need:				
Work or publication itself	10	6	10	7
Works or publications by author	5	3	3	2
Subject	147	89	134	89
Both work or publication + subject	1	1	1	1
Both author and subject	2	1	2	1
TOTAL	165	100	150	100
Types of subjects (as percentage of total sets):				
Works or publications as subject	8	5	8	5
Individuals as subjects	74	45	60	40
Geographical name	37	22	28	19
Chronological term	26	16	20	13
Discipline term	35	21	30	20
Other proper term	11	7	10	7
Other common term	85	52	71	47

[a] *Total of 165.*
[b] *Total of 150.*

common terms, the queries nonetheless also contain many of the category types that are distinctive to the humanities.

Personal tags We mentioned earlier that the presence of personal tags in the descriptions of individuals might be responsible for some of the distinctive features of the natural language statements. To test this hypothesis, the fifteen NLSs (9 percent of the total of 165) that contained personal tags were removed from the set of NLSs and the figures recomputed for the master set of categories. Table 4 presents the two computations in one. The figures in the two left-hand columns are for the entire set of 165 NLSs, and those in the right-hand columns are for the set of 150 NLSs without personal tags. (The fact that there are 150 NLSs dealing with searches for subjects, and the same number for NLSs without personal tags, is purely coincidental; the two sets are different.)

As the table shows, removing the natural language statements with personal tags changes the overall pattern very little. Differences between science and humanities statements do not appear to be due to the presence of personal tags.

Discipline terms As noted earlier, one of the surprises of this analysis was the large number of discipline-related terms that appeared in the natural language statements. The following discipline terms were mentioned once or more: architecture, art, art history, education, engineering, history, humanities, literature, music, philosophy, rhetoric, and science. These terms were generally not used as references to the fields as academic disciplines; rather, they most commonly referred to the artistic activity or the products of that activity that the scholar was studying, as in the following examples: "checking for articles on Nietzsche and music," "image of the tree in literature, art, science, of medieval and renaissance Europe," and "metropolitanism in American architecture." Another subset was represented by the common phrase, "history of": "any articles or book written about the problem and the history of imagination."

Others were part of longer, more specific phrases, such as "ephemeral art." Of the thirty-five discipline terms, eleven (or 31 percent) were of this type: part of a longer phrase that would make it a more specific term. The other twenty-four instances were freestanding.

These discipline terms are of particular interest for online searching, because one of the cardinal sins in online searching is considered to be the use of the name of an entire subject field in one's search. It is thought of as pointless to enter a term representing the subject of an entire database; one does not ask for "architecture" in an architecture database or "aluminum" in an aluminum database.

Even when the discipline and the database differ, as, for example, in indexing or searching for architecture in a literature database, search terms of this sort are assumed to be so broad as to be useless. Indexers seldom apply terms of such breadth. Consequently, requiring the presence of such terms in a search formulation will guarantee that all those relevant records that contain the other terms wanted by the searcher but lack discipline term indexing (the usual case) will fail to be retrieved. Free text matches might be made, but success with these is unpredictably dependent on the author's usage of discipline terms in title or abstract. Yet 23 percent of all the subject queries in this study contained a discipline term of some sort. (Problems would be less severe with the portion of discipline terms that are used in a longer, more specific phrase. As noted earlier, about a third of the discipline terms were of this type.) The results with respect to discipline terms raise questions about whether the design of online databases and their indexing schemes match humanities scholars' information needs as well as they might. The variety of types of discipline terms should be studied in greater depth.

Bibliographic features The third broad area of terms coded was bibliographic features. Two categories were coded: (1) preferences for bibliographic format (book, article, etc.) and (2) publication date or date ranges.

Only three natural language statements (or 2 percent) specified publication date or date range. Either this factor is unimportant for the scholars, or they specify such date ranges only later in their thinking during the search, perhaps when they are actually formulating the search statements.

Of the 165 natural language statements, forty-two (or 25 percent) mentioned some desired bibliographic form. On this bibliographic feature all forms mentioned in an NLS were also counted individually in order to identify frequencies of particular forms. Most of the forty-two NLSs mentioned articles: 20 percent (thirty-three NLSs) of the total 165 NLSs, with 9 percent of the total (fifteen NLSs) mentioning other forms, such as books and dissertations. (Figures add to more than forty-two NLSs because some statements mentioned more than one form.)

Often when "articles" were specified, the term seemed to have been used simply as the most convenient way to express the information need. At times, "articles" appeared to be used interchangeably with phrases like "literature on," "material on," and so forth. Since most of the databases the scholars searched include primarily citations to articles, the specification of "articles" was nearly meaningless. (Incidentally, when scholars used the word "literature" in the same way—that is, to indicate materials wanted—it was not counted as a discipline term.) All in all, specification of bibliographic features appears to be of only modest importance in the natural language statements.

Categories used infrequently During the several transformations of category sets, we experimented with other categories as well but dropped them because they applied to too few natural language statements in the sample to be of significance. However, a larger sample, or a different one, might find these categories worthwhile. Since all of the categories used in this study were developed with respect to a single sample, their validity should be tested and revised with other samples. In any case, there is some value in reporting that the candidate categories were not as productive as expected.

The study of movements is sometimes important in the humanities—movements among either the people studied ("nationalism") or those doing the studying ("the New Criticism"). In this study only four natural language statements contained such terms. Classes of works ("symphonies") and classes of creators ("painters") were considered as well but also produced few examples. The latter appeared primarily in personal tag lines. Finally, a very specific form of place is the name of a building or institution, and a very specific form of time is a named event. These, too, produced a smattering of occurrences. In later tabulations, when these trial categories were eliminated, the terms were tallied under the broader category to which they belonged.

Analysis of online search terms

Methodology

Introduction Because the analysis of scholars' natural language statements revealed an interesting array of subject categories that researchers into online searching discuss only infrequently, it seemed appropriate to analyze the actual search terms the scholars used in their DIALOG searches to see whether they also showed the same pattern of subject category assignments. Thus the approach taken here was to identify the scholars' search terms and assign them to the categories used in the analysis of natural language statements. Nonsubject categories of search, which proved in the first study to be a small part of the scholars' searching, were not analyzed. Otherwise, study conditions were the same.

The seemingly straightforward task of assigning subject search terms to categories proved surprisingly difficult. This task, which produced interesting results, also raised a number of methodological issues. The subsections below represent each of the major decisions made about how to analyze these data. Because we used a technical distinction between "type" and "token" (explained below), we will reserve the term "type" for that technical meaning and avoid other senses of the term in this section. Specifically, when referring to the categorization of terms, we will refer to categories of subject search terms rather than their types.

Definition of search term In our study of natural language statements it was easy to identify the natural language statement and difficult to distinguish terms within the statement. In this substudy, we encountered the opposite difficulty; the search term was easy to identify, but the query was not.

Let us define an online search statement as the searcher's input to the search system that begins immediately after a system prompt and ends with a carriage return or enter. In looking at any series of such search statements, it is difficult to identify with confidence the search formulation that a searcher intended to represent a search query. What the searcher thinks of as a single query may be expressed in one or several online search statements. For example, suppose a searcher combines two terms with a Boolean OR in one statement, then in the next statement uses a Boolean AND to combine the results of the previous statement with another term. Did the searcher have all three terms in mind all along and intend to create the final result through the two-step process, or did he or she originally intend to search only with the first statement, then revised it into another query after seeing the results of the first? In a study of the transaction logs recording the search statements it is impossible after the fact to be confident of what really constitutes a complete query as the searcher intended it.

On the other hand, it is easy to identify distinct terms within the formal, artificial syntax of the DIALOG command language. A search term is defined here as the string of characters bounded by the beginning or end of the search statement and by Boolean operators. This simple definition provided a basis for identifying and categorizing search terms. So in our analysis of online search terms the unit of analysis is the individual search term.

The search statement "Select Kandinsky AND German literature" contains two search terms: "Kandinsky" and "German literature." Thus, a search term may contain more than one word. It may also contain proximity operators, truncation, and prefix and suffix codes. For example, the entire phrase "urban (w) experience?/de" was considered a single term. Boolean operators within a proximity phrase or suffix-coded phrase were considered a part of the phrase and did not mark a separate term. For example, "urban(w)(experience OR milieu)" was treated as a single proximity phrase, and "(film? OR cinema? OR kino?)/ti" was treated as a single suffix-coded phrase. "Urban AND (experience OR milieu)," on the other hand, contains three distinct search terms.

Categories of subject terms Search terms were assigned to the same set of categories as in the analysis of natural language statements, with two exceptions. First, in categorizing geographical search terms it was often difficult to determine whether a term was used as an adjective or noun. Frequently, both categories were implied in truncated search terms such as "German?" which can refer to either "German" or "Germany." Thus all geographical name search terms were grouped into a single category. Second, in a few cases it was impossible to determine the proper classification of a search term or assign it to a single category. Those cases were placed in an "uncertain classification" category.

Types and tokens Linguistic analysis makes a fundamental distinction between a type count—a count of each distinctive term—and a token count—a count of each use of a term. Here is an invented example: If the total set of terms input by a scholar consisted of "Schoenberg," "Schoenberg," "Schoenberg," and "Europe," then this set of terms contains two types ("Schoenberg" and "Europe") and four tokens (all four terms listed above). The task of translating the type/token distinction into the online search environment proved to entail some subtle problems.

We decided to count types rather than tokens in order to eliminate repetitions and focus on distinct search terms. This study contained a number of repetitions of the sort shown in the example above. Users frequently revised search formulations, repeated a query during a later search session, or made errors of one sort or another and had to repeat the search formulation.

Our underlying interest was in identifying, categorizing, and counting the number of distinct search term types used by the searchers. Since our focus was on search terms (as defined in an earlier subsection), not just terms in general, it was desirable to identify and count search term types, not just general term types. It is necessary to understand the implications of this choice to interpret the resulting data.

Two kinds of grammar were operative for the terms studied here: that of standard English and that of the artificial DIALOG command language. To illustrate the consequences of this situation, consider the truncated and untruncated search terms "art?" and "art." In DIALOG's command language, the presence of the question mark produces a search that matches on a very different set of terms in the database indexes than would be the case without the question mark. "Art" can match only with "art," while "art?" can match with "artist," "article," and the like. To search using one and then the other version of the term would be a perfectly rational act for a searcher. In standard English, however, "art" and "art?" might be considered two tokens of the same type. To reflect the online searching environment, our count treated the two terms as distinct search term types rather than as two tokens of one type.

Many of DIALOG's grammatical features produce similar results. With proximity operators, for example, the "(w)" operator requires that words in matching terms be adjacent and in the same order as in the proximity phrase, while the "(n)" operator requires that they be adjacent but in either order. The addition of numbers within these operators, such as "(2w)," allows the search words to be matched even if that number of words, or fewer, intervene. Just as with truncation, the difference between "history

(1w) philosophy" and "history (1n) philosophy" allows for different possible matches with the database in the actual search, even though the words themselves are the same. In the case of the "(n)" operator, the above example will match with history of philosophy as well as philosophy of history, whereas the "(w)" version will match only with the former. Since the family of matching phrases allowed by each proximity operator combination is different, the two phrases were defined as distinct types.

Similarly, the presence or absence of suffix codes was considered to mark distinct types rather than merely tokens. For example, "architecture," "architecture/de," and "architecture/de,ti" were considered to be three search term types. Suffix codes, by definition, stand for fields that DIALOG has determined are subject related and are therefore contained within the "Basic Index." As a variant in search technique, a searcher may specify that a term be searched on only some of the several subject-related fields (generally, descriptor, title, abstract, and one or two others, depending on the database). In that case, the searcher enters a suffix code for the desired subject field(s).

Another kind of DIALOG code, the prefix code, such as "AU =" for author, almost always refers to fields other than subject. Since in this study the use of subject-related prefix codes was very small, all uses of prefix codes were excluded.

Combination terms "Search term" has been operationally defined as the string of characters bounded by Boolean operators and the beginning or end of a search statement. However, many scholars combined into one phrase—sometimes using correct DIALOG syntax and sometimes not—search terms containing distinct categories as defined in this study. Here are three examples: "Beethoven (n) pastorale," "ferrarese (w) painters (w) fifteenth century," and "sex? germ? (lit? OR cult?) mod?" Interpreting these phrases for online searching is problematic. The first two phrases are technically correct in their use of the DIALOG syntax, while the third phrase uses incorrect syntax. For a variety of reasons, the first two phrases would fail in a search in many databases, but not necessarily all. In particular, *Art Literature International* (that is, *RILA*) uses long, compound headings of this sort, which might match if the search terms were worded correctly.

As a practical matter, and whether the search formulation is correct or incorrect, it is impossible to classify these long phrases in one or the other of the categories used here. "Beethoven (w) pastorale" is made up of a person and a work, two elements that need to be classified independently. Unlike our analysis of natural language statements, in which distinct concepts

TABLE 5. *Search term tokens and types*

	NUMBER
Search term tokens	2,467
Search term types:	
Single formal	931
Combination formal	137
TOTAL FORMAL SEARCH TERMS	1,068

NOTE: *All figures (both tokens and types) include search terms without affixes and with suffixes. Prefix-coded terms are excluded.*

were difficult to identify with confidence, we found that once the search term as a whole was demarcated by the DIALOG syntax, it was generally not difficult to identify such distinct concepts within the larger term. We therefore decided to identify distinct terms *within* search terms when we believed that the searcher clearly intended distinct concepts.

In the analysis these are called "combination terms." They were tallied independently of the single search terms, as well as combined with them in a total figure. To distinguish the two senses of "search term," we called one the "formal" sense (bounded by Boolean operators, etc.), and the other the "informal" sense (distinct concepts identified by the study analysts).

It was usually relatively easy to distinguish between the separate informal search terms in a combination term because different categories of terms, such as composer (individual) and symphony (work) were stated. Sometimes, however, terms of the same category occurred together in the same phrase. An example is "form gestalt," which represents both the English and German words for the same term. We decided to categorize the phrase "form gestalt" as a combination term, with both terms categorized as "other common" terms. "Weimar? (2n) america?" is another example of combining two concepts of the same category, by using two distinct geographical names in the same phrase. In such cases, these were also considered "combination terms" and were given their own categories: "other common + other common" and "geographical + geographical," respectively.

Results

As noted earlier, a distinction was made in the tally between (1) the "formal" definition of search term as the string of characters bounded by Boolean

TABLE 6. *Single search terms input by scholars*

SUBJECT SEARCH TERM CATEGORIES	*N*	%
Works or publications as subject	44	4.7
Individuals	351	37.7
Geographical name	59	6.3
Date or date range	11	1.2
Period	16	1.7
Time modifier	0	0
Discipline term	23	2.5
Other proper term	50	5.4
Other common term	362	38.9
Uncertain classification	15	1.6
TOTAL	931	100.0

operators and the beginning or ending of the search statement and (2) the "informal" definition of search term as a conceptually distinct term that fell within a term as demarcated by the formal definition. Where only one conceptually distinct (informal) term was identified within the formal term, formal and informal term were by definition identical. We considered a "single term" to be a formal term that contains just one informal term, and a "combination term" to be a formal term that contains two or more informal terms. Therefore, by definition, the sum of all single and combination terms (the whole combination, not its constituent terms) is equal to the sum of all formal terms.

Looking first at the results for formal terms, table 5 provides figures for actual search terms input by scholars (tokens), as well as total single and combination formal terms (types) input by the scholars. The ratio of types to tokens is 43 percent.

Table 6 presents totals for categories of all single formal terms, and table 7 presents totals for categories of all combination formal terms. Here are some examples of the search terms included in the count in table 6: "portrait painting" is an other common term; "alexandria" is a geographical name; "newton? (w) isaac" is the name of an individual; "renaissance" is the name of a period. Here are examples of combination terms included in the count for table 7: "unesco, bibliotheca" (other proper + other common); "mantegna (w) miniature" (individual + other common); "humanities (w) (method? OR methodology)(w)(comparison usa europe)" (discipline + three other common + two geographical name).

TABLE 7. *Combination search terms input by scholars*

SUBJECT SEARCH TERM CATEGORIES	*N*	%
Other common + other common	20	14.6
Other common + date	2	1.5
Other common + discipline	15	10.9
Other common + geographical	28	20.4
Other common + individual	12	8.6
Other common + period	5	3.6
Other common + other proper	3	2.2
Discipline + discipline	7	5.1
Discipline + geographical	7	5.1
Discipline + period	2	1.5
Geographical + geographical	2	1.5
Geographical + period	2	1.5
Geographical + other proper	3	2.2
Individual + individual	6	4.4
Individual + period	1	.7
Individual + other proper	1	.7
Individual + work	7	5.1
Period + other proper	2	1.5
Other proper + other proper	1	.7
Other common + date + discipline	2	1.5
Other common + date + geographical	2	1.5
Other common + discipline + geographical	2	1.5
Discipline + geographical + period	3	2.2
Date + discipline + geographical	1	.7
Other common + discipline + geographical + period	1	.7
TOTAL	137	99.9

NOTE: *Percentages add to less than 100 because of rounding error.*

In table 7, for simplicity's sake, combinations containing more than two instances of the same category, of which there were few, were grouped with the dual category. For example, the term "(film OR cinema? OR kino?)/ti" was grouped in the "other common + other common" category.

TABLE 8. *Total informal search terms (both single and combination) input by scholars*

SUBJECT SEARCH TERM CATEGORIES	*N*	%
Works or publications as subject	51	4.1
Individuals	384	30.8
Geographical name	114	9.1
Date or date range	18	1.4
Period	32	2.6
Time modifier	0	0
Discipline	70	5.6
Other proper term	62	5.0
Other common term	500	40.1
Uncertain classification	15	1.2
TOTAL	1,246	99.9[a]

[a] *Adds to less than 100 percent because of rounding error.*

As noted earlier, the formal definition of search term sometimes did not match with what the scholars clearly intended as search terms. All the combination terms counted in table 7 occurred because the scholars grouped several (informal) search terms together within a single (formal) search term as normally defined. Many, if not most, of these combinations were ineffectual, since they represented incorrect use of DIALOG's search features.

Since the informal terms often appeared to represent the distinct concepts the searchers originally intended, it is valuable to see how many distinct informal terms of each subject category they used, whether in single or in combination search terms. Table 8 presents these data.

We may illustrate the coding for table 8 using the last example listed above for table 7 ("humanities (w) method? . . . ," etc.). That single formal combination term contained the following informal search terms: one discipline, three other common, and two geographical name. So for that term, a total of six term counts will be added to the tally in table 8: one discipline, three other common, and two geographical.

All terms in table 8 are types, rather than tokens, as in tables 6 and 7, with one exception: while multiple uses, that is, tokens, of formal combination search terms as a whole were purged and reduced to types, individual informal terms within combinations were not checked for duplication with single terms or terms in other combinations.

TABLE 9. *Percentages of NLS appearances and informal search terms in each category*

TYPE OF SUBJECT SEARCH TERM	PERCENTAGE OF ALL NLS "APPEARANCES"[a]	PERCENTAGE OF ALL INFORMAL SEARCH TERMS[b]
Works or publications as subject	3	4
Individuals as subject	27	31
Geographical name	13	9
Chronological term	9	4
Discipline term	13	6
Other proper term	4	5
Other common term	31	40
Uncertain classification	0	1
TOTAL	100	100

NOTE: *Data in column 1 are extracted from column 1 of table 2; data in column 2 are extracted from table 8.*

[a] *Total of 276.*

[b] *Total of 1,246.*

All three tables reveal roughly the same pattern as found in our analysis of natural language statements. Humanities scholars use a wide variety of categories of terms in their searches, just as in their natural language statements. In table 8, 1 percent of the terms fall into the "uncertain" classification, 40 percent into "other common," the most common category by far in Saracevic's science queries, and all the rest (59 percent) fall into some other category.

At this point it would be desirable to compare the two sets of results more closely. Do the scholars' natural language statements contain the same percentages of the various categories as their search terms? The results of these two substudies cannot, in fact, be properly compared, for reasons discussed earlier. The unit of analysis in the first substudy was the entire natural language statement; figures are for numbers of statements containing one or more instances of a category, that is, "appearances." The unit of analysis in the second substudy was the individual search term. Thus no comparison between the two substudies can be exact, and no comparative statistical analyses have been done.

However, while keeping the above caveats in mind, it is possible to get an approximate sense of the relationship between the data in the two substudies. Table 8 presents the percentage of all informal search terms in each category. By reanalyzing the data in table 2, we produced a unit of analysis closer to that used in table 8.

If we add the total appearances listed in the first column of table 2 (rather than the total NLSs), we may compare percentages of total appearances for each category in the first substudy against percentages of total informal search terms in the second substudy. These figures are only roughly equivalent, but nonetheless revealing.

Table 9 compares percentages of each category in the two substudies. (The sum from col. 1 in table 2, or 276, is based on total geographical and total chronological terms, not on individual categories, such as "date or date range," within these broad categories. In table 9, all kinds of geographical and all kinds of chronological terms are combined.)

If we again keep in mind how approximate this comparison is, the two kinds of units—natural language statement appearances and informal search terms—exhibit a similar pattern. However, some of the term categories that are most distinctive to humanities researchers' work—geographical, chronological, and discipline terms—show up less frequently in informal search terms than in natural language statements. Is this because researchers do not need these terms for an effective search, or because they cannot find ways to express these aspects of their information needs effectively in the online environment?

On the other hand, searches on individual names, another category of term more common for humanities searchers, increased when researchers turned their natural language statements into search terms. Did our method of counting terms in NLSs ("appearances" of one or more terms in a category) undercount the number of distinct instances of the "individual" category in the NLSs? Or are individual names easier to express in searching than other subject categories needed by these scholars? In the interviews, to be detailed in a later report, one scholar mentioned that he had resorted to more searching on individual names because he had difficulty formulating searches for other kinds of subject terms. Was his experience typical? These questions will be discussed in the next section.

Discussion and conclusions

Twenty-two Getty Visiting Scholars produced the data that are analyzed in this first report of the Getty Online Searching Project. Most of the total of 165 natural language statements (NLSs) describing projected online searches constituted searches for subjects, as opposed to searches for works or authors. Searches on individuals as subjects were very popular; 45 percent of the NLSs mentioned them. Geographical names, terms referring to dates and historical periods, and discipline terms were popular as well.

Categories of terms used were compared to those of a major study conducted by Saracevic and Kantor and funded by the National Science Foundation (Saracevic & Kantor, 1988). The NSF study analyzed the results of online searching done largely on social and natural science queries. Virtually no terms for works as subjects, individuals as subjects, or geographical, chronological, or discipline terms appeared in that study; the terms that overwhelmingly predominated were what our study identified as "other common" terms: 100 percent of the science statements contained these terms, while only 57 percent of the subject natural language statements in this study contained them. On the other hand, combining all term categories besides other common, only 18 percent of the science queries used these other terms categories, while 84 percent of the Getty subject natural language statements did. The contrasts between the kinds of terms found in the two studies were very marked overall.

The second part of the study, which categorized scholars' online search terms, produced roughly similar results. Of all the distinct informal term types the scholars used in their searches, 59 percent were other than other common.

Owing to various factors operating in the data analysis, results of the two substudies were not easily compared. Means were found, however, to produce a rough comparison between the percentages of subject categories appearing in searchers' natural language statements and search terms, respectively. It was found that search terms referred to relatively more individuals and other common terms than natural language statements did, and to relatively fewer geographical, chronological, and discipline terms.

The fact that both substudies showed a pattern of frequent use of terms besides other common complements the work of Stephen Wiberley (1983, 1988), who studied the access points in encyclopedias, dictionaries, and abstracting and indexing services in the humanities. Although he used different categories, which makes direct comparison impossible, Wiberley also found a high incidence of personal names and other proper nouns. In the studies described in this article, natural language statements of information needs and online search terms also contained many such terms. Thus, the information needs of scholars (analyzed here) and the design of the resources intended to meet them (analyzed in Wiberley's articles) appear to match in a general way.

However, Wiberley also found that the percentages of proper terms varied widely from one humanities field to another. The percentage of what he called "singular proper" terms, that is, terms that name "a person or creative work whose existence in space and time has been ascertained" (Wiberley, 1988, p. 26), ranged from 24 percent in a philosophy index to 93 percent in an English literature index. If we combine our categories of "works" and "individuals," the result should be fairly close to Wiberley's singular proper. Individuals, as defined here, can include mythical and fictional characters, but otherwise the term is similar to Wiberley's.

A combined total of 50 percent of the natural language statements referred to both works and individuals as subjects. Thirty-five percent of the informal search terms constituted the same combination of categories. The humanities abstracting and indexing service that most closely reflected the interests of the Getty scholars was *RILA* (*International Repertory of the Literature of Art*). In Wiberley's count, 48 percent of the terms in *RILA* were singular proper, a figure that fits well with the figures in the substudies here. (Note that logically there is no reason that the figures should necessarily be close. We would expect that a hypothetical perfect arts and humanities information database would have many of both kinds of terms, but, to meet researchers' information needs perfectly, it might be necessary, for example, for it to contain a proportionally larger set of singular proper names than would appear in scholars' queries.)

In her dissertation, Helen Tibbo (1989) asked a sample of historians what kinds of information they would like to have in abstracts of historical materials. The four highest-ranked categories of information they mentioned, and the percentages of respondents requesting that category, were as follows: (1) specific dates and time span indicators (100 percent), (2) names of geopolitical units (100 percent), (3) names of individuals and/or groups (96 percent), (4) main topic or subject of work (92 percent) (Tibbo, 1989, p. 540).

The equivalent of other common does not appear on this list until fourth place. Elsewhere Tibbo concludes that "the facets of time, place, and specific topic" are used by historians "to delimit their research, classify their literature, and organize college curricula" (Tibbo, 1989, p. 591). Historians, too, want many of the kinds of terms our group of scholars wanted. Thus, three large empirical studies—Wiberley's, Tibbo's, and ours—all point to the importance of terms besides other common to humanities researchers.

This contrast between other common and all other term categories is an important one, for several reasons. The other common terms that overwhelmingly predominated in the science queries of Saracevic and Kantor's NSF study in fact constitute the heart of what are considered "subject" terms in much writing on thesauri and search vocabulary. Modern principles of thesaurus development received their chief impetus after

World War II, when science and technology libraries and indexing services needed more detailed and technically accurate indexing than previous systems had made possible. Apparently, however, the predominance of science and engineering in the early days of thesaurus theory development led to a heavy emphasis on other common terms and a corresponding underemphasis of non–other-common terms.

The problems are not limited to thesauri. Even the *Library of Congress Subject Headings*, which predates the development of modern thesaurus principles, and which makes detailed provision for geographical, period, and form subdivisions, uses some period subdivisions only to subdivide extra-large files. Where there are many entries under a main geographical heading the subdivisions are used; where there are few, they are not. With such unpredictable application—that is, unpredictable for the searcher—it is impossible to make reliable use of these non–other-common subdivisions in online searching. Yet, clearly, for the humanities scholar, meaningful online searches can generally be carried out only through use of both non–other-common and other common terms, with the emphasis on the former.

We suspect—although it would take another study of a different kind to determine this—that the relative underemphasis of non–other-common terms in thesauri and database indexing makes the use of such terms more difficult for online searchers. This may explain why scholars in the study made relatively less frequent use of several categories of non–other-common terms in search terms, in comparison to their use in natural language statements.

The Getty Art History Information Program (AHIP), now the sponsor of the bilingual *Bibliography of the History of Art* (*BHA*), the successor to *RILA*, is also currently conducting several projects relating to non–other-common terms. The results of these projects should contribute to solving some of the humanities scholar's problems highlighted by our research. Some facets in the *Art and Architecture Thesaurus*, such as the "Styles and Periods" facet, codify certain kinds of non–other-common terms that are relatively neglected in thesauri. *The Getty Union List of Artist Names*, scheduled for release in 1992, is a database of several hundred thousand artists' names, drawn from nine Getty projects, that clusters together the often voluminous name variations referring to a single artist, both during the artist's lifetime and afterward.

Geographical names are problematic because geographical jurisdictions and names change through time. Coverage of geographical names is frequently limited in thesauri and, consequently, in indexing. The Getty's *Thesaurus of Art-Historical Place Names* (*TAP*), a hierarchical database of place names now in preparation, should make possible complete and consistent retrieval of geographical materials. Ultimately, we hope to see catalogers and information users able to draw upon these various resources in a linked "virtual" database of descriptive terms, all developed to a uniform standard accepted throughout the world of scholarship.

The results have other implications for online searching as well. Research, practical searching advice, and the teaching of online searching all give relatively short shrift to the non–other-common terms identified here. Searching on names is discussed relatively little (for example, Everett & Pilachowski, 1986). As noted earlier, one scholar commented that he often searched on names because he found searching on other kinds of terms more difficult. The question of whether other scholars do the same merits more research.

These results concerning the use of chronological and discipline terms have particularly interesting implications for online searching. Dates and date ranges need to be represented in certain ways in bibliographic records to permit effective online retrieval. *Historical Abstracts*, a database in which dates are of obvious critical importance, experimented for several years before finding ways of coding dates that allowed flexible retrieval, that is, that allowed searches to be made using both stringent and loose requirements ("high-precision" and "high-recall" searches, respectively). Proper handling of dates for effective online retrieval is not obvious and is not yet widely understood. (See also discussion in Bates, 1992.) Dates should be considered another category of term used in thesauri, and thesauri should instruct indexers in how to represent chronological terms effectively. Only in that way will good date indexing make effective online retrieval possible.

The frequent use of discipline terms, in the natural language statements in particular, creates a paradox for the indexing and searching of online databases. Because these terms are very broad, they are normally seen as poor candidates for use in either indexing or retrieval. Yet these terms are often meaningful for humanities scholars. The scholars in the Getty project used them in nearly a quarter of their natural language statements. Only 6 percent of their search terms were discipline terms, however. Had they "learned their lesson" and discovered how ineffectual such terms are for online retrieval, given current indexing practice?

The sample of discipline terms used in this study was too small for confident identification of all the senses in which they are used, but scholars in various fields clearly use them in a variety of ways. Perhaps it would be possible to introduce these discipline terms into database indexing and retrieval by providing a special classification that indicated the various special senses in which humanities scholars use them. Discipline terms would thus have more specific meanings and provide more precise retrieval for humanities searchers. This question merits additional study.

Finally, our results point to another online searching issue in the humanities. Table 7 provides statistics on 137 search terms that were combinations of (informal) terms within single (formal) terms. In other words, within a single phrase bounded by the beginning or end of the statement and by Boolean operators, the scholars combined terms in all the different ways displayed in the table—twenty-five ways in all.

Many of these combinations would not be effective in a search; they are examples of the novice searcher still trying to master a not-so-simple formal command language syntax. But the large number and complex variety of these combinations shows how often humanities scholars need to combine the various term categories identified in the study. And the combinations could be fruitful: combining several broad terms (for example, discipline + date/period + geography) could lead to a narrow search. The point we have been trying to make here has no bearing on whether the scholars in the Getty study were expert online searchers: we are arguing that if the necessary indexing terms have been omitted from the databases through ignorance of their potential usefulness to the scholarly process, even expert searchers will be hampered in their searching efforts.

In the NSF study data for the natural sciences (table 3), on the other hand, every statement contained other common terms and only one statement contained one or more other proper terms. Thus, in the NSF natural science data, only two of the combinations listed in table 7 of this report were possible even in principle: other common + other common and other proper + other common.

The need, therefore, of humanities scholars to combine terms from a wide variety of distinct categories suggests that searching in the humanities may be inherently more complex than in the sciences. Yet there are many problems associated with developing good search formulations even in the sciences. Thus, the particular problems associated with effective searching in the humanities merit far more attention than they have been given to date.

This study has revealed the distinctive characteristics and needs of humanities scholars with respect to both thesaurus features and online searching. Thesauri developed for the humanities need to give attention to the non–other-common categories of terms identified here (works and individuals as subjects, geographical, chronological, discipline, and other proper terms), in addition to the other common terms. Indexers and searchers alike need to be able to draw upon the controlled vocabulary of each category consistently and easily.

Because of the very large number of terms needed for good coverage of individuals, works, and geographical names, it is probably unrealistic to include terms from all these categories, as well as others, in a single thesaurus for a given humanities discipline. However, term development in these various categories should be coordinated within a common framework or philosophy, and the results made available to indexers and searchers in a convenient form.

Likewise, training for searchers and help screens and other online aids for searchers, as well as search capabilities in online systems, need to be designed to take full cognizance of the unique characteristics of online searching in the humanities. Geographical, chronological, and discipline terms in the online environment hold particular promise for improvement.

To summarize, this study and other recent studies demonstrate that information seeking, vocabulary, and online searching in the humanities have many unique features that have been given relatively short shrift in theory and practice in library and information science. It is understandable that library and information science research has attempted to follow standard scientific practice by trying to generalize principles of thesaurus development and online searching across all disciplines. It may be, however, that in doing so some critical distinctions between disciplines have been overlooked, which has meant that certain groups of users have been underserved. The next stage of development in the theory of the field may therefore be to give closer attention to the unique features that differentiate subject literatures and disciplines—an intriguing prospect.

REFERENCES

Bates, M.J. (1992). Implications of the subject subdivisions conference: The shift in online catalog design. In M.O. Conway (Ed.), *The future of subdivisions in the Library of Congress Subject Headings system* (pp. 92–98). Report from the 1991 Subject Subdivisions Conference sponsored by the Library of Congress. Washington, DC: Library of Congress Cataloging Distribution Service.

Boyles, J.C. (1987). Bibliographic databases for the art researcher: Developments, problems and proposals. *Art Documentation, 6*(1), 9–12.

Boyles, J.C. (1988). The end user and the art librarian. *Art Libraries Journal, 13*, 27–31.

Broadbent, E. (1986). A study of humanities faculty library information seeking behavior. *Cataloging & Classification Quarterly, 6*(3), 23–36.

Case, D.O. (1991). The collection and use of information by some American historians: A study of motives and methods. *The Library Quarterly 61*(1), 61–82.

Corkill, C., & Mann, M. (1978). *Information needs in the humanities: Two postal surveys. CRUS Occasional Paper No. 2.* (BLR & DD Report No. 5455). Sheffield, England: Centre for Research on User Studies.

Crawford, D. (1986). Meeting scholarly information needs in an automated environment: A humanist's perspective. *College & Research Libraries, 47*(6), 569–574.

Everett, D., & Pilachowski, D.M. (1986). What's in a name? Looking for people online—humanities. *Datebase, 9*(5), 26–34.

Gould, C.C. (1988). *Information needs in the humanities: An assessment*. Stanford, CA: Research Libraries Group.

Guest, S.S. (1987). The use of bibliographical tools by humanities faculty at the State University of New York at Albany. *Reference Librarian, 7*(18), 157–172.

Horner, J., & Thirlwall, D. (1988). Online searching and the university researcher. *Journal of Academic Librarianship, 14*(4), 225–230.

Hurych, J. (1986). After Bath: Scientists, social scientists, and humanists in the context of online searching. *Journal of Academic Librarianship, 12*(3), 158–165.

International Organization for Standardization. (1986). *Documentation, guidelines for the establishment and development of monolingual thesauri* (2nd ed.). (International Standard ISO 2788). Geneva: International Organization for Standardization.

Katzen, M. (1986). The application of computers in the humanities: A view from Britain. *Information Processing & Management, 22*(3), 259–267.

Krausse, S.C., & Etchingham, J.B., Jr. (1986). The humanist and computer-assisted library research. *Computers and the humanities, 20*(2), 87–96.

Lehmann, S., & Renfro, P. (1991). Humanists and electronic information services: Acceptance and resistance. *College & Research Libraries, 52*(5), 409–413.

Loughridge, B. (1989). Information technology, the humanities and the library. *Journal of Information Science, 15*(45), 277–286.

Lowry, A., & Stuveras, J. (1987). *Scholarship in the electronic age: A selected bibliography on research and communication in the humanities and social sciences*. Washington, DC: Council on Library Resources.

Mackesy, E.M. (1982). A perspective on secondary access services in the humanities. *Journal of the American Society for Information Science, 33*(3), 146–151.

Rahtz, S. (Ed.). (1987). *Information technology in the humanities: Tools, techniques and applications.* Chichester, England: Ellis Horwood/John Wiley.

Ross, J.E. (1987). Artists and poets online: Issues in cataloging and retrieval. *Cataloging & Classification Quarterly , 7*(3), 91–104.

Ruiz, D., & Meyer, D.E. (1990). End-user selection of databases—part III: Social science/arts & humanities. *Database, 13*(5), 59–64.

Saracevic, T., & Kantor, P. (1988). A study of information seeking and retrieving: II. Users, questions, and effectiveness. *Journal of the American Society for Information Science, 39*(3), 177–196.

Schmitt, M. (Ed.). (1988). *Object, image, and inquiry: The art historian at work.* Santa Monica, CA: The Getty Art History Information Program.

Schmitt, M. (1990). Alas, the failure to communicate: Thoughts on the symbiosis of scholars, information managers and systems experts. *Art Documentation, 9*(3), 137–138.

Stam, D.C. (1984). How art historians look for information. *Art Documentation, 3*(4), 117–119.

Stam, D.C., & Giral, A. (Eds.). (1988). Linking art objects and art information. *Library Trends, 37*(2), 117–264.

Stern, P. (1988). Online in the humanities: Problems and possibilities. *Journal of Academic Librarianship, 14*(3), 161–164.

Stielow, F., & Tibbo, H. (1988). The negative search, online reference and the humanities: A critical essay in library literature. *RQ, 27*(3), 358–365.

Stone, S. (1982). Humanities scholars: Information needs and uses. *Journal of Documentation, 38*(4), 292–312.

Taylor, R.S. (1968). Question-negotiation and information seeking in libraries. *College & Research Libraries, 29*(3), 178–194.

Tibbo, H.R. (1989). *Abstracts, online searching, and the humanities: An analysis of the structure and content of abstracts of historical discourse.* (Doctoral dissertation). University of Maryland, College Park, MD.

Walker, G. (1988). Online searching in the humanities: Implications for end-users and intermediaries. *Proceedings of the 12th International Online Information Meeting* (pp. 401–412). Oxford, England: Learned Information.

Walker, G. (1990). Searching the humanities: Subject overlap and search vocabulary. *Database, 13*(5), 35–46.

Walker, G., & Atkinson, S.D. (1991). Information access in the humanities: Perils and pitfalls. *Library Hi Tech, 9*(1), 23–34.

Weinberg, B.H. (1988). Why indexing fails the researcher. *The Indexer, 16*(1), 3–6.

Wiberley, S.E., Jr. (1983). Subject access in the humanities and the precision of the humanist's vocabulary. *The Library Quarterly, 53*(4), 420–433.

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

Wiberley, S.E., Jr. (1991). Habits of humanists: Scholarly behavior and new information technologies. *Library Hi Tech, 9*(1), 17–21.

Wiberley, S.E., Jr., & Jones, W.G. (1989). Patterns of information seeking in the humanities. *College & Research Libraries, 50*(6), 638–645.

Woo, J. (1988). The Online Avery Index End-User Pilot Project: Final report. *Information Technology and Libraries, 7*, 223–229.