

Multiple Dimensions of Search Results

Tamar Sadeh*

Abstract - The library community—librarians and developers of software for libraries alike—has invested a great deal in organizing and describing scholarly information and in developing searching techniques and query interfaces to enable library users to discover that information. Until recently, little effort had gone into the display and management of search results: these were typically presented in a one-dimensional list, regardless of the number of results in the list. A method of presenting results in a way that is meaningful and useful to searchers is now emerging and beginning to show its impact on scholarly research. By incorporating clustering—the grouping of search results according to the similarity of words and phrases—or faceted categorization—the grouping of results on the basis of predefined, structured metadata that is available for scholarly materials—information systems can render the linear display of results into a multidimensional structure and improve the findability of the content that searchers are seeking.

Keywords – faceted categorization, clustering, search interface

I. INTRODUCTION

Now that online searching has reached a certain maturity, the information retrieval community is realizing that for information systems to find the items that are relevant to a user's query is not enough. The new challenge is for the systems to present the results of a query in such a way that users will be able to easily locate the results that fulfill their needs. Among the first industries to encourage the development of features that increase a user's probability of finding relevant items were companies in e-commerce, because the "findability" of products is crucial for business survival. Since 2000, search engines such as those from [Endeca Technologies](#) and [Mercado](#) have been enabling e-commerce systems to delve into metadata and bring to the surface information that makes a multidimensional presentation of results possible. Among the scholarly information systems that are joining this emerging trend are [Scirus](#) from Elsevier, the [WorldCat](#)[®] catalog from OCLC, [AquaBrowser](#)[®] Library from [Medialab Solutions](#), the [Primo](#)[®] and [MetaLib](#)[®] systems from [Ex Libris](#), and library catalogs that deploy the Endeca engine (such as the [North Carolina State University catalog](#)). Such information systems now incorporate clustering—a feature that enables them to group search results according to the similarity of words and phrases—or faceted categorization—the grouping of results

on the basis of predefined, structured metadata that is available for scholarly materials. By presenting results in this multidimensional fashion, information systems help users understand the content of a given result list and home in on the most relevant materials.

II. SEARCHING IN SCHOLARLY INFORMATION SYSTEMS

The organization, description, and storage of scholarly data have always been the target of much attention. In particular, libraries and information providers have invested in creating rich metadata that describes scholarly data and in designing query interfaces that include a multiplicity of options. The perception of the designers of scholarly information systems has been that users take advantage of the complex query interfaces to accurately specify their information needs, thus enabling the systems to easily match the users' specifications with the metadata and find the appropriate results.

For example, the basic query interface of the [HOLLIS Catalog](#) at Harvard University provides a dropdown list for selecting one option out of the almost 30 that are offered for Search Type, such as *Keywords anywhere*, *Title beginning with...*, and *Medical subject beginning with...* In addition, the HOLLIS Catalog provides an Expanded Search option (Figure 1), in which the user can combine several fields with Boolean operators, request that the system regard the search terms as separate words or as phrases, and limit the search by using various criteria such as language, location, format, and date range. The Command Search option offered by the HOLLIS Catalog enables expert users to type their queries in a command line.

Information providers in the scholarly arena build their interfaces in a similar way, providing both simple and advanced query interfaces. For example, the Elsevier [ScienceDirect](#)[®] interface offers three query levels: Quick Search, Advanced Search, and Expert Search. In the Quick Search interface, users can specify one or more words from an article's title, abstract, keywords, or author's name; part or all of a journal or book title; the volume; the issue; and a page number. The Advanced Search feature offers more search options, including a subject and date range, and enables the user to combine a few queries to form a new search. The Expert Search feature enables users to enter search terms along with Boolean operators.

* Ex Libris Ltd., Malha Technological Park, Jerusalem 91481, Israel. E-mail: tamar.sadeh@exlibrisgroup.com.

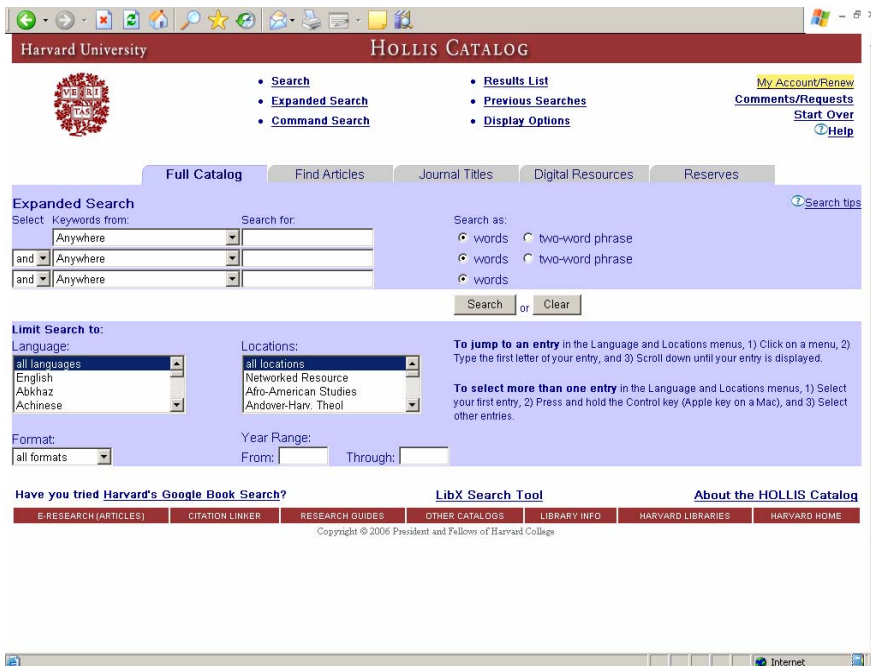


Figure 1. Expanded Search in Harvard University's HOLLIS Catalog

Whereas in the past, users took advantage of at least some of the options that query interfaces offered, recent surveys point out that searches have become simpler: searchers enter fewer words per query, hardly ever use Boolean operators, and rarely use any of the more advanced search options. According to K. Markey, "people enter a few short search statements into online IR systems. Generally, their queries bear two to four words. Boolean operators are uncommon. Boolean operators are even few in number in the searches of end users who receive training in system use. End users rarely use advanced system features and when they do, they are quite likely to use them incorrectly. Although research findings demonstrate that end users are not conducting sophisticated online searches, the vast majority are satisfied with their searches. In fact, percentages of users who express satisfaction with the results of their searches reach into the high seventies and beyond" [1] (see also [2]). Other surveys indicate that academics are drawn to simpler interfaces like those of [Google](#) and [Google Scholar](#) and are willing to trade off the accuracy and trustworthiness of the traditional information resources against quick, friendly, and easy-to-use interfaces [3], [4]. If we take into account that today's searchers became familiar with the interfaces of Web search engines before encountering scholarly query interfaces, it is no wonder that such searchers are intimidated by the complexity of the latter. Furthermore, many searchers are not subject experts and cannot define their information needs in a precise way, especially when they conduct an exploratory search about a topic.

Despite the effort that libraries and information providers have invested in search engines and query interfaces, not much progress has been made in the display and management of search results—although e-commerce sites such as [eBay](#) and [Amazon.com](#) (Figure 2) have implemented new ways of handling results to improve their products' "findability" (defined by Peter Morville as "The quality of being locatable or navigable...The degree to which a particular object is easy to discover or locate...The degree to which a system or environment supports navigation and retrieval" [5]). In the scholarly environment, search results are typically presented in a one-dimensional list, regardless of the number of items in the list.

However, with the huge amount of data that is available today (and growing exponentially) and the indifference shown by typical users toward the phrasing of their search queries, many searches in the scholarly environment generate a large number of results. A purely linear presentation of so many results is inadequate, despite the use of sophisticated relevance-ranking algorithms to prioritize the result list. When applied to scholarly materials, these algorithms, though enabling systems to display results in an order that users prefer, are questionable in that they lack the context in which the query was defined and hence cannot assist the system in tailoring the presentation of the results to the specific person and need. Furthermore, relevance-ranking algorithms alone cannot bridge the gap between a user's intended query and the way in which the user phrases it.



Figure 2. Findability enhanced at Amazon.com by the grouping of results by topic

In addition, information-seeking patterns that searchers have evolved in other contexts, such as in a quest for non-scholarly information on the Web, lead them to focus on the first results in the list, regardless of its length. In a query for a specific, “known” item (such as a particular article), this behavior may prove adequate because the searcher is likely to have entered enough information (for instance, the article’s title and author) for the system to find the exact match, which will then be displayed at the top of the result list. However, when a user conducts an exploratory search to find items relevant to a particular topic, the first results might not be applicable; the reason could be that the query terms were too broad or that they differed from the metadata terms used by the system. For example, an inexperienced searcher might look for *brain and mind* rather than choose a more professional term, such as *neuropsychology* or *cognitive neuroscience*, which are the terms that are likely to be found in the metadata. Moreover, query terms can be interpreted in more than one way, especially when used for searches in information resources that are not targeted at a specific discipline. For example, a searcher might type the acronym *ABC*, which is an abbreviation for the basic steps of cardiopulmonary resuscitation (*airway, breathing, circulation*), a measurement of the sustainable harvest level of a fishery (*acceptable biological catch*), the name of a light car developed in the 1920s by the British company ABC Motors, and many other phrases.

Some systems try to overcome the challenges of exploratory searches by providing lists for browsing. Systems such as library catalogues enable the searcher to browse through subject headings and drill down in a controlled hierarchy of subjects until the exact subject is found (Figure 3). At this point, the searcher can examine all items that the library classified under the specific subject. This type of search process requires familiarity with the hierarchy of the subject headings. The process can be long and cumbersome and can lead to a dead end when a subject heading has no items at all. Furthermore, browsing through a hierarchy requires a single selection at every stage. That is, once the user selects one

subject heading, all the others are excluded. A user who is interested in civil wars in both North America and South America, for example, will probably have to repeat the browsing process because the subject headings, such as those from the [Library of Congress Authorities](#) list, are either too general (*Civil war*) or too specific (*Civil war Brazil*). Another obstacle to the novice user is that the organizational logic of subject headings is not evident; the user might not know that the best way to find materials about the American Civil War is to start with *United States*, then choose *History*, and then choose *Civil War, 1861-1865*—rather than start with *Civil war*.

Because users are accustomed to Web search engines, most do not opt for browsing through subject headings. Rather, such users are likely to conduct an exploratory search by entering one or more terms that describe their information needs and then trying to make the most of the result list. To help such users, designers of scholarly information systems are starting to pay more attention to postsearch processing.

There are several ways in which a system can increase the findability of items once the initial result list is displayed. For example:

- If a search yields no results or too few results because of a poor query term or a misspelling, the system can suggest alternative searches (“did you mean?”), which might enable the user to obtain the desired results.
- If a search yields a large number of results, the system can enable the user to drill down to subsets of the list (Figure 4, Figure 5).
- Regardless of the length of the list, the system can analyze the results and on the basis of the analysis, suggest that the user drop the initial query and search the whole collection again for one of the topics associated with the current result list.



Figure 3. Beginning of the list of subject headings at the Library of Congress

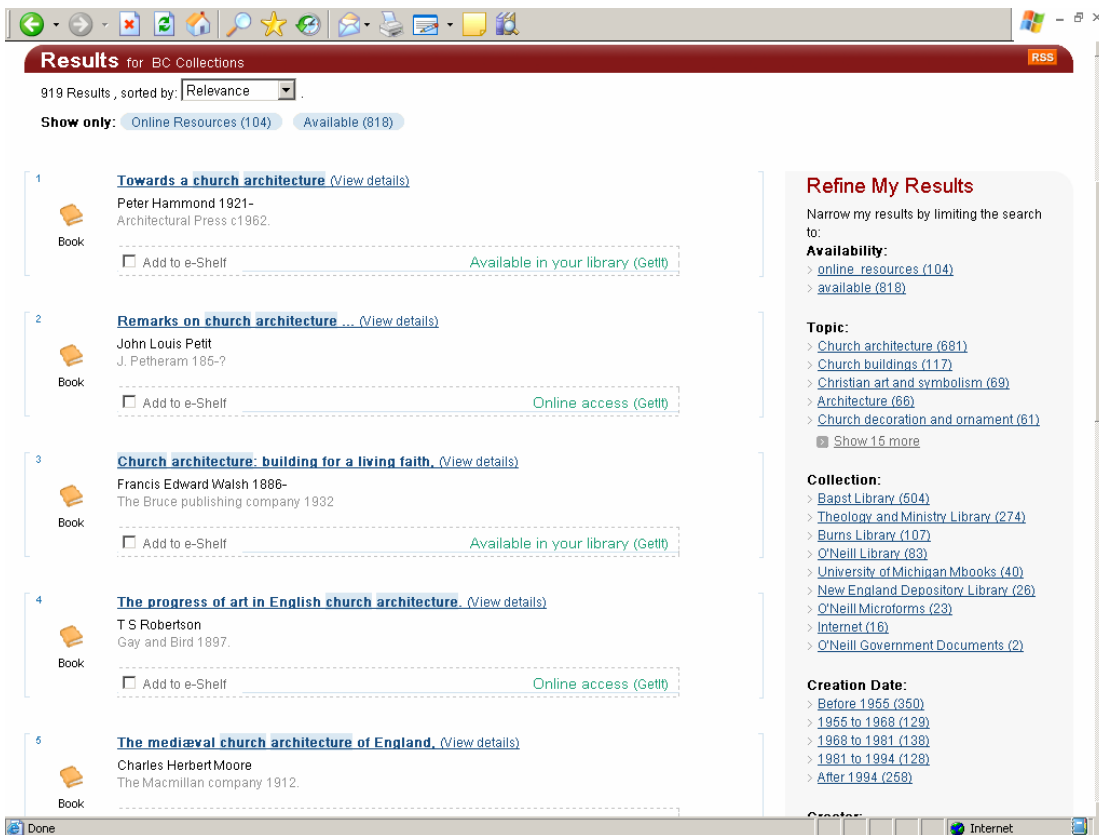


Figure 4. Result list presented in Super Sleuth, at Boston College, enabling users to drill down through subsets of the list

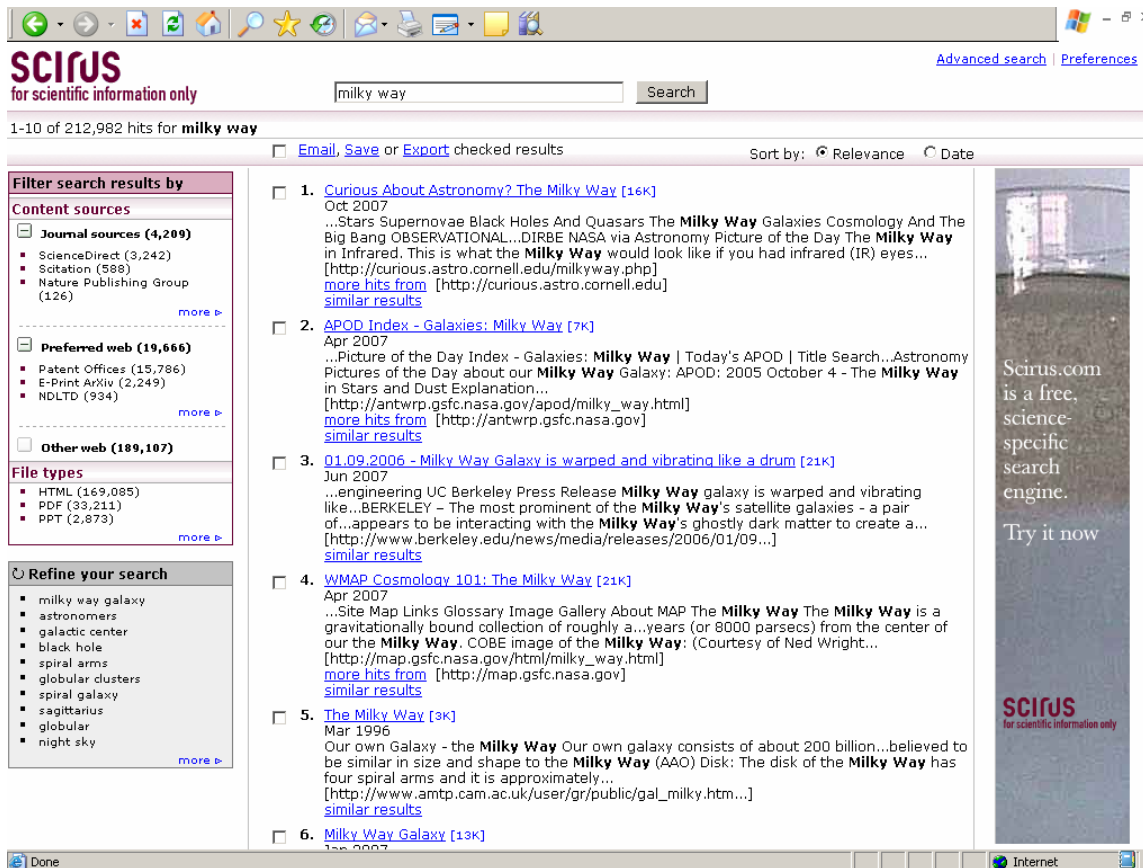


Figure 5. Results presented in Scirus, enabling users to refine their search by selecting a topic

What these options have in common is the notion that the result list itself is a tool that can be exploited. One of the ways in which the system can exploit the list is by analyzing it and presenting it as a multidimensional structure; users can then examine it from various angles, gaining a better understanding of its content and the way it applies to their query. A multidimensional structure also helps users easily identify and discard those results that are not relevant.

III. CREATING A MULTIDIMENSIONAL STRUCTURE FOR A RESULT LIST

Each item on a result list of scholarly materials is associated with metadata, a set of library-defined attributes, which can include subject, author, type (such as a book, article, journal, or image), genre (such as biography, bibliography, fiction, or film and video adaptation), language, year of publication, and the collection in which the item is stored. Some attributes can have only one value, such as year of publication, while other attributes can have several values, such as number of authors or number of subjects. The values for the subject attribute can be derived from an authoritative list such as the Library of Congress Subject Headings vocabulary or the National Institutes of Health Medical Subject Headings (MeSH®) vocabulary. A scholarly information system can create multiple dimensions from a result list by presenting the list of attributes alongside the result list and, for each such attribute, a list of the values that are shared by the records in that group (for example, one of the attributes could be the type of item, with the values “book”, “journal”, “article”, and so on) (Figure

6). Each time the user selects a value, the result list is redisplayed to show all the items that share that value. In this manner, the system enables the user to look at the results from different angles, each time focusing on a specific characteristic of the result list. Furthermore, the list of attributes and their values provides a summary of the results: at a glance, the user can see which topics characterize the result list, which types of items the list contains, in which languages they are written, and so on.

The system has two ways of identifying the values of the attributes that will constitute the multiple dimensions of a result list. In one method, typically referred to as *faceted categorization* or *faceted browsing*, the system groups the results on the basis of one attribute at a time—extracted from library-defined, structured metadata—such as subject or date of publication. In the second method, typically referred to as *clustering*, the system looks through numerous metadata fields of each record and extracts phrases that are repeated in multiple records. Let’s look at a case in which the system locates the phrase “chronic fatigue” in the subject field in record 1, the phrase “Anxiety, fatigue, and privation” in the abstract field in record 2, and the phrase “adrenal fatigue” in the description in record 3. In this case, the system might identify the word “fatigue” as the common thread among the three records, put them in the same group, and label it “fatigue,” thus indicating to the user that all records under this label are related to the fatigue.

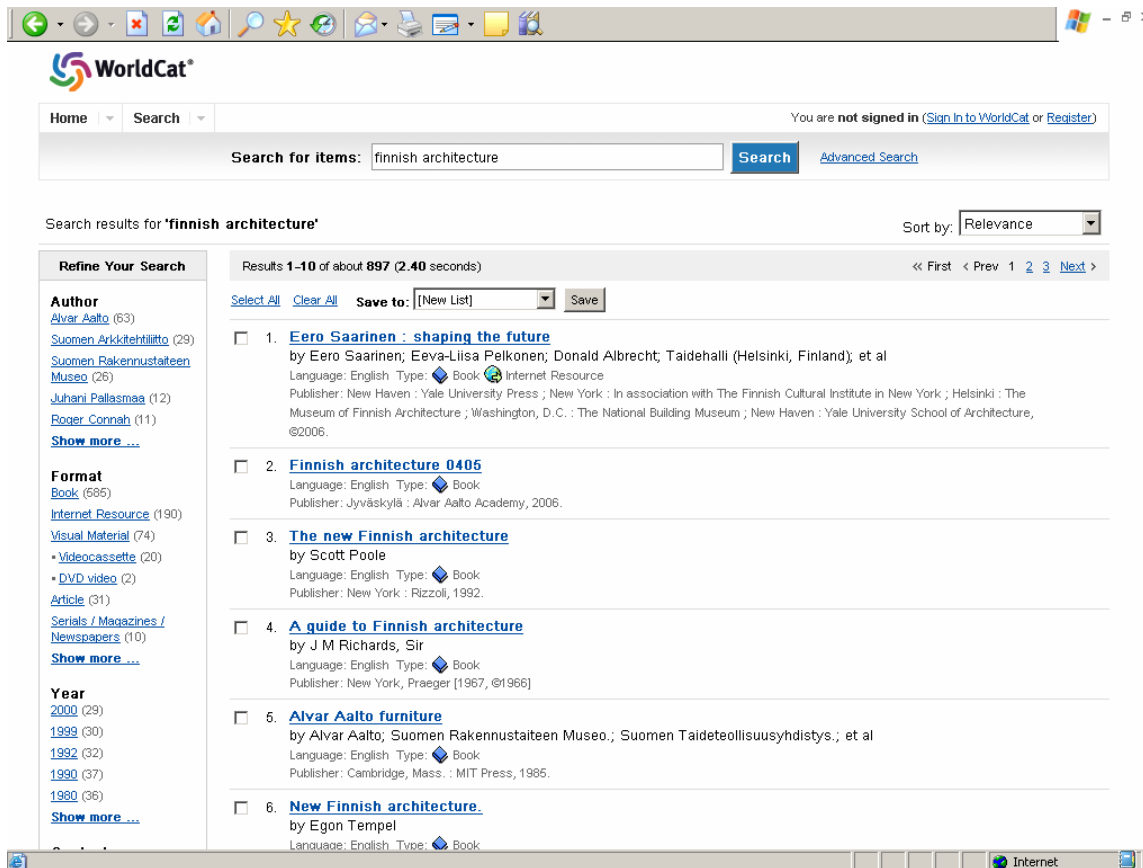


Figure 6. Result list presented along with its groups in OCLC WorldCat

The way in which a scholarly information system displays the grouped items and provides access to subsets of the result list is the same for the two underlying technologies that are used to create the groups. Moreover, both technologies can be deployed for the same result list. For example, the system might have used clustering to group records by topic but faceted categorization to group records by material type, language, publication date, and genre. A searcher can select any of the options presented: when looking at the language group, the searcher can choose the option *Chinese*, for example, and narrow down the list to see only the items written in Chinese; when looking at the topic group, the searcher can choose any of the topics and narrow down the list to see only the items that are relevant to that topic. The selection process is repetitive; that is, every time the user selects an option, a new result list is displayed. On the basis of this list, the system updates the groupings and enables the user to select a new option. The user can continue selecting options for any of the attributes until a concise, manageable result list is attained.

To understand the benefit of multiple dimensions in a result list, let us think of a searcher who is looking for materials related to English poetry of the eighteenth century but doesn't have a specific item in mind. A novice searcher may start with a general query, such as *English poetry*, without specifying any restrictions of the metadata fields in which the query terms appear. Such a query is likely to yield many thousands of results, the first of which are not necessarily relevant to the user. For example, when one searches the catalogue of the Library of Congress for *English poetry*, the system returns 10,000 results—the upper limit on the result list—with the first ones as follows:

- On whistler mountain* by Martyn Crucefix (1994)
- Loving letters: an Islamic alphabet* by Riad Nourallah (1995)
- Fayre formez of the Pearl poet* by Sandra Pierson Prior (1996)
- Sad piper/poems* by Omar Tarin (1994)
- Stumbling dance* edited by Rupert M. Loydell (1994)
- Poems of three generations* by Stafford Cripps et al. (1977)
- Poetry India: voices for the future* edited by H.K. Kaul (1993)
- Muppet babies giant book of rhymes* illustrated by Tom Brannon (1994)
- S'adi-e-Hind: Hasan Dehlavi* by Nargis Jahan (1989)
- Twinkle, twinkle, little star: a lullaby book with lights and music* illustrated by Jannat Messenger (1988)

Clearly, none of the results makes sense for the specific need. When the user rephrases the query to make it more specific—for example, *English poetry eighteenth century*¹—the result list is of the same size (10,000 results) but some of the records displayed at the top of the list seem slightly more relevant:

- William Collins and eighteenth-century English poetry* by Richard Wendorf (1981)
- Poetry of the landscape and the night: two eighteenth-century traditions* (1970)
- Byron's "Don Juan" and the eighteenth-century English novel* by András Horn (1976)

¹ Subject headings from the Library of Congress Subject Headings classification are shown here with the capitalization used in that classification scheme.

Eighteenth-century English poetry: the annotated anthology edited by Nalini Jain and John Richardson (1994)

New Oxford book of eighteenth century verse chosen and edited by Roger Lonsdale (1994)

Silence and sound: theories of poetics from the eighteenth century by Richard Bradford (1992)

Points in eighteenth-century verse; a bibliographer's and collector's scrapbook (1972)

Language of natural description in eighteenth-century poetry (1966)

Landscape, liberty, and authority: poetry, criticism, and politics from Thomson to Wordsworth by Tim Fulford (1996)

Literary transmission and authority: Dryden and other writers by Jennifer Brady et al. (1993)

Still, when scanning this list, the user will never be able to find all, or even most, of the items that she thinks will fulfill her needs. A user who is more familiar with library searching may try a query with *English poetry eighteenth century* in the subject field; however, this search yields no results, because the correct subject heading, according to the Library of Congress Subject Headings classification, is *English poetry 18th century*. A search for the broader subject—*English poetry*—shows that 2,108 books have the subject *English poetry* and 377 more books have the same subject but as a subject heading for children. Then there is a list of 21 subject headings that start with *English poetry 18th century*, such as *English poetry 18th century History and criticism*, *English poetry 18th century History and criticism Bibliography*, and *English poetry 18th century History and criticism Congresses*. Each such subject is a gateway to one or more items, and it is unclear whether items that show up when the user chooses one subject will also appear when the user chooses another subject; would a book that is located under *English poetry 18th century History and criticism* also appear when one chooses the subject *English poetry 18th century*?

A system that analyzes the result list and offers multiple views of it creates a completely different experience for the same user with the same query. For instance, a search for *English poetry* in the [Smart Search](#) system of the University of Iowa libraries, an implementation of the Ex Libris Primo[®] technology,² also yields many results—19,129—but the system provides the searcher with tools to narrow down the list by examining various aspects of it. For instance, the user can choose the topic *English poetry 18th century* and focus on the 270 items that share this topic.

The information that is likely to be the most helpful to a user who is trying to target relevant items is the list of topics related to the results. By presenting these topics, the system enables the user to understand what kinds of items appear on the list—even way down the list—and concentrate on those topics that are relevant. The list of topics can be organized according to the number of applicable items, so the user can see right away which topics are more dominant. The user can decide to focus on *English poetry 18th century* but can also choose another subject that is closer to the area of research, such as *Literature, Comparative; Romances, English; Literary form*; or *Monologue*. Because the process is repetitive, the

user can first focus on one topic, say, *English poetry 18th century*, and then, once the list of topics that apply only to the records related to the query *English Poetry* and the topic *English poetry 18th century* is displayed, the user can narrow down the choices even farther by selecting another topic. What is more, the list of topics can show relationships of which the user may not have been aware; for example, when the same query—*English poetry*—is submitted in [Super Sleuth, Boston College libraries'](#) implementation of Primo, the topic *Political satire, English* comes up. Choosing that topic brings up the book *Poetry and politics under the Stuarts*, which, although not classified under *English poetry 18th century*, may be relevant to the user's research.

The list of topics can help in many other cases. When the user enters an author name that is common, the topic list can help the user locate the works of the specific author. When the term itself is used in multiple disciplines, the list of topics enables the user to exclude items from non-related disciplines. An example is the term *mercury*, which can apply to astronomy, chemistry, engineering, literature, and even music.

Grouping the results according to the topics that the result list represents is only one of the many ways in which a system can present results to users. A system can group results based on other criteria, as well. If we look at the Smart Search system (Figure 7), we see that it groups results according to several criteria, including:

- Availability—enabling a user to separate out the online items, for example
- Type of item—enabling the user to divide the list into books, journals, audiovisuals, and so on
- Collection in which the item is stored
- Language of the item
- Publication date
- Genre
- Library of Congress classification headings

² For information about this technology, see <http://www.exlibrisgroup.com/category/PrimoOverview>.

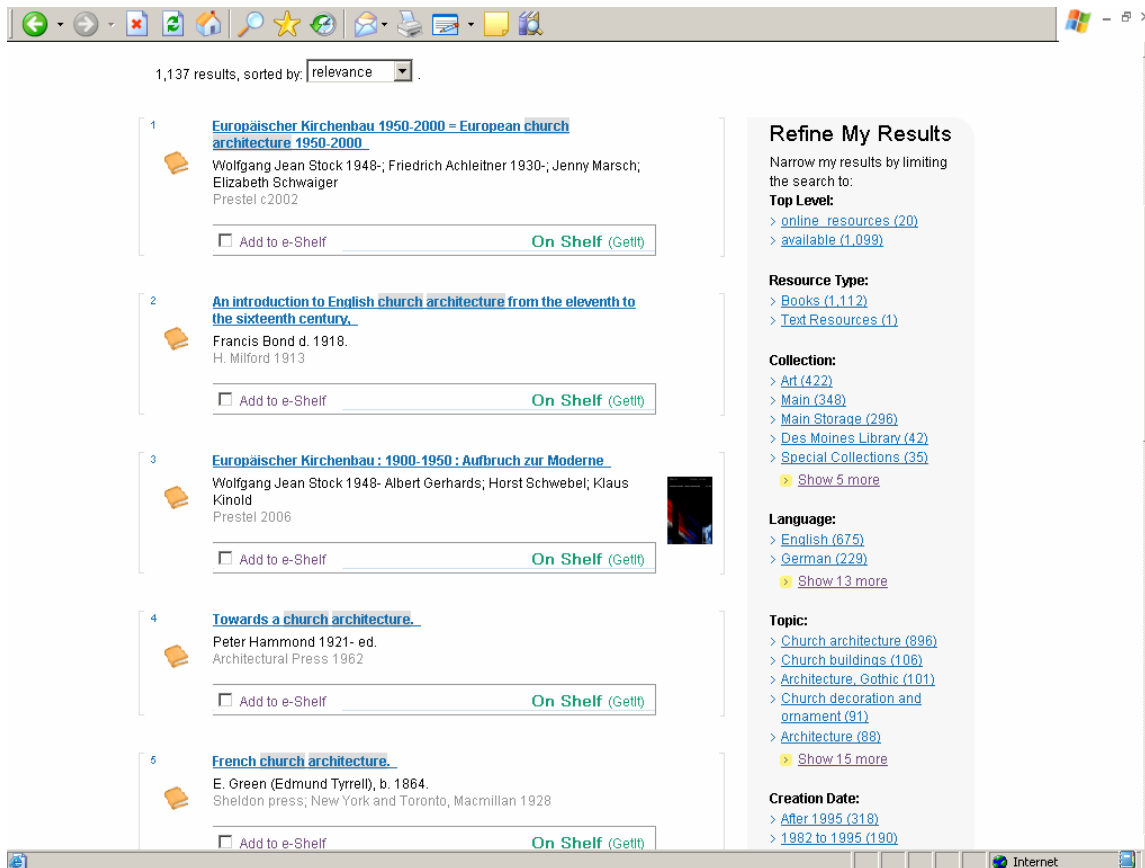


Figure 7. Faceted categorization, shown in Smart Search at the University of Iowa

Because the system identifies the possible groupings on the fly on the basis of the specific result list, all the groups that the system offers apply to items in the result list. There are no dead ends: the user will not see any reference to groups that contain no items. For example, if the list does not contain bibliographies, the genre group will not include a reference to bibliographies. Furthermore, the number of items associated with each attribute option is displayed in the window (Figure 8).

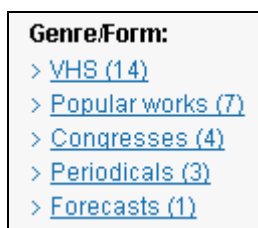


Figure 8. List of genres characterizing items in a specific result list

Such a structure of the result list enables the searcher to combine various criteria to quickly narrow down the list. A searcher who is using Super Sleuth to find new books about the Romantic Movement in eighteenth-century English poetry may start with the query *English poetry 18th century*, which yields 600 results (Figure 9). Looking at the various attributes that characterize the items on the list, the user can click the topic *Romanticism*, thus limiting the list to 61 results. Examining the attributes presented for the 61 results, the searcher can select the books with a recent publication date, limiting the list to 21 results—a size that is manageable. In only two selections after the original query, the searcher has managed to focus on the relevant items.

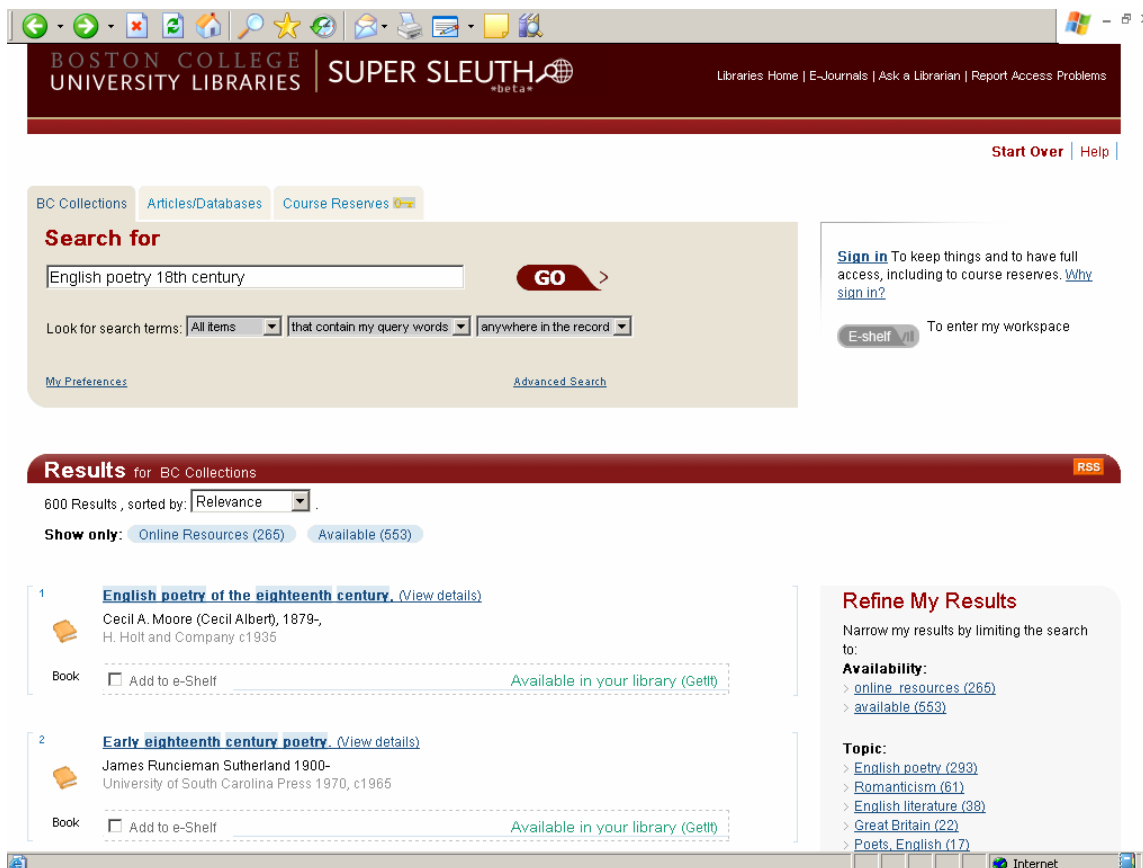


Figure 9. Results for the query *English poetry 18th century* presented in Super Sleuth

IV. FROM METADATA TO FACETS AND CLUSTERS

Usability studies have shown the benefits that can be achieved, from the user-experience point of view, when the system creates a multidimensional result list [6], [7], [8], [9]. However, the success of the multidimensional structure depends on the system's ability to gather meaningful metadata from the records in the result list and to group the records on the basis of similarities in the metadata. The task of gathering metadata is not trivial; metadata carries with it certain challenges. Do the relevant records have metadata? What is its breadth? How accurate is it? To what metadata schema does it adhere?

Some providers of information resources invest more effort than others in creating or assigning comprehensive and coherent metadata. Furthermore, the homogeneity of a collection is crucial for the grouping of similar items; a set of records that share the same attributes, and the same values for those attributes, enables the system to group the results in a meaningful way. A set of records that do not share the same attributes or whose attributes' values adhere to different authoritative lists are less likely to enable the system to create high-quality groupings. For example, if some of the results have *genre* as an attribute and some do not, either the latter results will be completely absent from the groups that the system creates for the attribute *genre* or, if the system defines a catchall *genre* group for those cases in which records have no *genre* attribute, the results will be found in such a group (which might be labeled *other*, for instance). In either of these two cases, a user who selects the genre *bibliographies* will miss all the bibliographies that were not assigned a *genre* at all. Another example relates to the uniformity of the attribute values: if some records are assigned a subject from one list of

subject headings, such as the Library of Congress headings, and other records are assigned a subject from another list, such as the MeSH headings, the grouping of records by subject becomes challenging.

A library catalogue can serve as an example of a collection that is fairly homogeneous in terms of the metadata assigned to each item and therefore enables the system to present a well-defined list of attributes. In such a catalogue, the adherence of author names to an authoritative list of authors permits excellent faceted categorization by author: all references to a specific author, say, Jane Austen, will have the exact same form, and hence when a user searches for *women novels 19th century*, the system will group all Jane Austen novels under one author name—for example, *Austen, Jane, 1775-1817*. If the values of the attributes are not well defined, the author name could be phrased in more than one way (*Austen, Jane*; *Austen, J.*; *Jane Austen*; and so forth). As we can see, the importance of the quality of the metadata is not limited to the search; the quality is also crucial for the processing of the result list.

Very few systems can provide optimal metadata. In many cases, metadata was defined over long periods of time and is therefore inconsistent, reflecting different trends. In other instances, the items originate from various collections and hence their metadata adheres to different rules. New items are usually catalogued in a more standard and uniform way: after a long tradition of cataloguing in-house, today's libraries and information providers typically obtain cataloguing information from authoritative national or international resources. Nevertheless, libraries and information providers often add metadata or modify some of it. When a system tries to extract attribute values, the inaccuracies and inconsistencies of the metadata become apparent.

Furthermore, many of the systems that enable users to search for scholarly information provide access to more than one information resource, each of which might deploy its own metadata schema in a way that applies to that resource. For example, with the Super Sleuth system at Boston College, users can search for materials originating from the university's library catalogue, course materials created at Boston College, the institution's digital collections, and external resources, such as [MBooks](#) (the digitized book collection of the University of Michigan). Since the system must match attribute values that originate from each of these resources, it needs to adjust these values so that they all conform to one form. Such adjustments, referred to as "normalization," may include the removal of punctuation and extra spaces and the changing of all the letters to the same case. However, normalization can only help when the differences between attribute values are minor; hence, heterogeneous collections, which may differ considerably in the metadata that is defined for their items, pose many challenges to faceted categorization.

The advantage of using metadata, if, indeed, it is of high quality, is the coherence and uniformity of the list of attributes and their values. As a result, the user can expect a standard set of attributes for every result list and knows that the absence of a specific attribute signifies that it is not applicable to the list. For example, if the result list contains no items written in Chinese, Chinese will not be listed under the *Language* attribute; and if all the items on the list are in English, the *Language* attribute will not appear at all. Furthermore, all values of a specific attribute are always of the same type; for instance, only values such as *English, French, Dutch, and Chinese* can be listed under the attribute *Language*. The predictability of the attributes is reassuring, on the one hand, but limiting, on the other hand: only groupings that were predefined will be presented to the user.

Clustering is another approach to grouping. As explained earlier, clustering refers to the process of grouping items on the basis of similarities between words or phrases across metadata fields, or, as Hagedorn et al. put it, "taking words and phrases that make up metadata records and gathering them together into semantically meaningful groupings. For instance, a record about the feeding and care of cats can be grouped with a record about the feeding and care of hamsters" [6]. In addition to finding similarities, the clustering engine labels them; that is, it assigns a textual description to each similarity.

The clustering approach proves to be more effective than faceted categorization when the available metadata is incomplete or inconsistent, such as in a search across collections. In Web searches, clustering is the only method that can work because of a lack of structured metadata assigned to Web pages.

The clustering engine of [Vivísimo](#) is demonstrated through [Clusty](#), an application implementing the Vivísimo Velocity search technology over the Web.³ When a user searches for *Nobel prize winners*, for example, the Clusty application displays the first results (228 results out of "at least 550,200 retrieved," according to the site), and clusters them on the fly to form, initially, 10 groups (Figure 10).

The labels of the groups—for example, *university, physics, and chemistry*, demonstrate the power as well as the weakness of clustering. On the one hand, the clustering engine is not restricted to a specific metadata field but matches words and phrases across fields; hence, it can identify the commonalities between the records and bring to light concepts that were not necessarily expected yet make sense. On the flip side, the list of values is neither coherent nor complete. In this example, *literature, economics, chemistry, and physics* refer to areas for which a Nobel Prize is awarded, whereas *university* connects universities that publicize their graduates who are Nobel Prize winners; and *photographs* might have been picked by the system because pages about Nobel Prize winners also often point to photographs or photo galleries. Users may find this assortment of groupings confusing. Another issue worth mentioning is that the clustering engine dealt with only the first 228 records. The limit is imposed, in this case, because the clustering is performed on the fly: fetching over half a million records from the Web, analyzing them, determining which terms characterize them, and grouping the results according to these terms would prove costly in respect to time and resources. Some would argue that a list of 228 results is more than enough for the user to process, particularly because the Web search engines probably already sorted the results by relevance before returning them to Clusty. However, one of the great benefits of clustering, as with faceted categorization, is that the groupings provide an overview of the whole result list. This benefit is lost if grouping applies to only a very small subset of the result list. As Hearst notes: "The disadvantages of clustering include...[clusters'] lack of predictability, their conflation of many dimensions simultaneously, the difficulty of labeling the groups (Clusty.com's top-level labels are among the best implementations) and the counterintuitiveness of cluster subhierarchies" [10].

Faceted categorization relies on attributes that are specified for each item. When the system is grouping items, it identifies facets without considering other items' attributes or any particular result list. Clustering, on the other hand, is applied to an entire list of items, and the labels that the clustering engine generates for part of the list might differ from those that it generates for the list as a whole. Furthermore, the clustering engine can build on one set of words and phrases describing a specific item when it appears on one result list and another set when the item appears on another result list. For example, an article about the works of the architect Antoni Gaudí can be grouped with articles about churches if the initial query concerned churches, yielding many church-related results; but the same article will be grouped with articles about ceramic tiles if the initial query happened to be about tiling in Barcelona. These groupings are, of course, not necessarily related to any of the subjects that a cataloguer might have assigned to the Gaudí article; the article's similarity to other articles may have surfaced through the title, the abstract, or another field. (The Vivísimo technology uses only the title and the abstract to detect similarities.)

³ For information about this technology, see <http://vivísimo.com/html/products>.

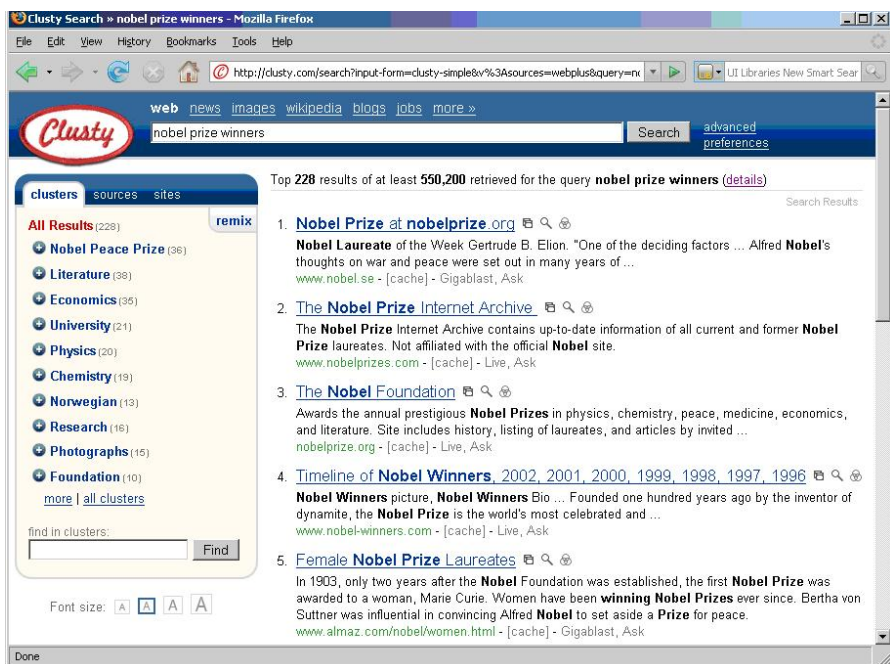


Figure 10. Clustered results presented in Clusty

Vivísimo’s clustering engine has been incorporated in the MetaLib[®] metasearch system from Ex Libris. Using MetaLib, a searcher can submit a query to heterogeneous information resources. The system broadcasts the query to each resource and returns the results to the user in a uniform format [11]. When presenting the results to the user, MetaLib employs both faceted categorization and clustering to make the result

list multidimensional: values that are extracted from metadata fields that tend to be uniform across information resources—author, date, journal title, and information resource (database)—are used for faceted categorization, and the Vivísimo clustering technology is deployed to create groups of results by topic, because subject headings tend to differ across information resources (Figure 11).

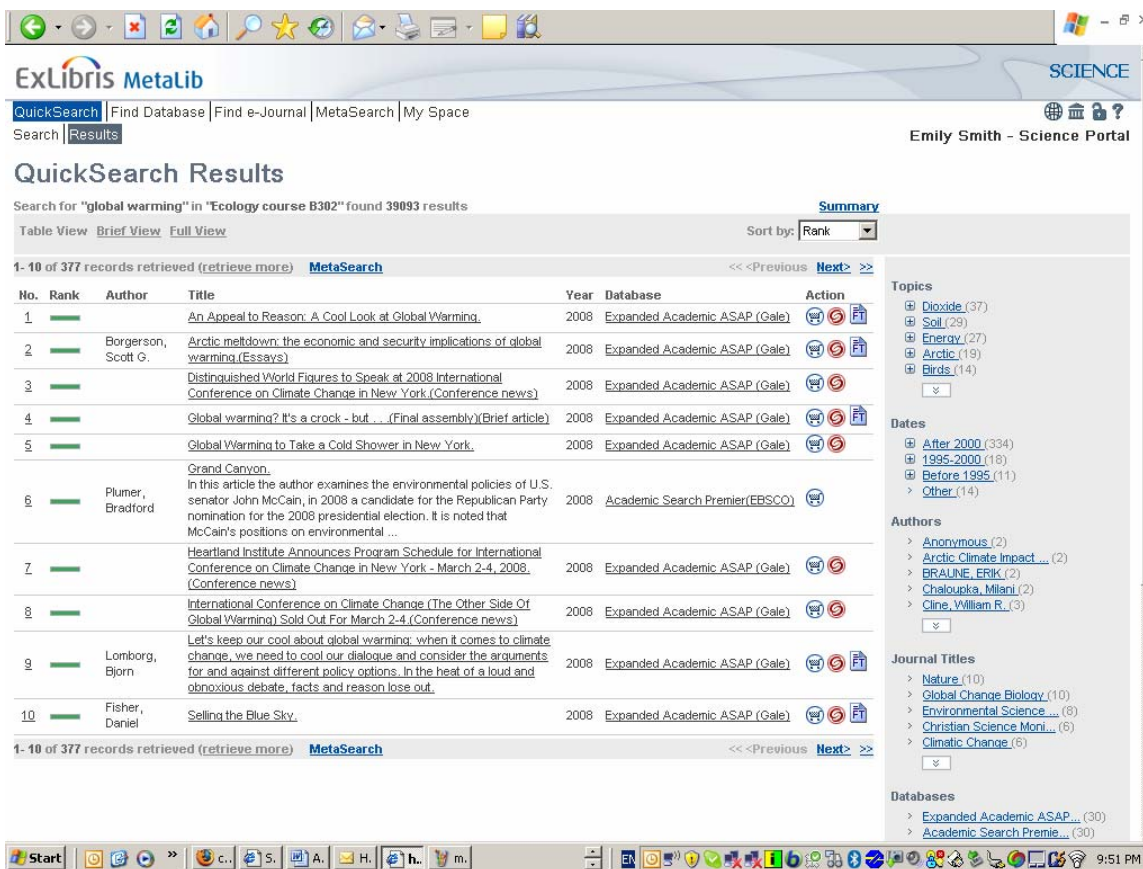


Figure 11: Clustering and faceted categories in a MetaLib result list

To demonstrate how metadata enrichment and automated classification deployed in a large, heterogeneous collection

affect users’ ability to search the collection, a team from the University of Michigan and the University of California,

Irvine, carried out a project that was described in a January 2007 report [6]. The team used a topic modeling algorithm to automatically group a collection of 2.5 million records, originating from 388 OAIster⁴ repositories, into 500 clusters. Out of the 500 clusters, 352 were usable in the sense that the words that were picked up by the system as characterizing the group seemed to belong together or had a strong subject affiliation with each other. Staff from the Digital Library Production Service and the Scholarly Publishing Office of the University of Michigan University Library then manually examined the clusters and labeled them with a word or phrase that would define or describe each group. The next step was to map the labels to a classification system, embed the labels and the classification categories in the data, and build a user interface layer that would take advantage of the cluster labels.

According to the team's report, the users did benefit from the automatically generated clusters:

. The clustering methods we used enabled us to improve the findability of appropriate materials by providing additional subject categories for use in searching and browsing. These new terms were then used as selections within the search interface, on a new browse page, and as a way to filter the search results. We quickly realized that the real power of including new subject terms was on the search results page. To reiterate the benefits, we provided:

- narrowing and/or expanding the results using the "Results by Topic" facet,
- the freedom and context of viewing records in individual categories,
- the discovery of different results by choosing a different category, and
- clarification of vague or broad search queries [6].

The project demonstrates a novel approach, whereby clusters are generated without regard to any particular query or result list. Because the clusters are created in advance and their labels are embedded in the metadata records, the system is able to group records in a manner that resembles faceted categorization; that is, the system uses the cluster labels in the same way it uses any other metadata field. Despite the reported gain for end users, the effort of creating the clusters and embedding them in the metadata was rather intensive. Furthermore, such a process does not seem applicable in all cases, particularly when the metadata cannot be preprocessed.

V. CONCLUSIONS

Faceted classification and clustering are two technologies that enable scholarly information systems to present search results in a way that makes sense to users. These features render a purely linear representation of a result list—as long as it may be—into a multidimensional structure where items are intelligibly grouped by various attributes or aspects. In addition to gaining an immediate grasp of the contents of the result list, users can easily focus on items that are more relevant to them by selecting any of the groups that the system presents.

Usability studies have confirmed that searchers welcome the multidimensional display. Also, according to Morville,

faceted classification coheres with the way people understand information and provides the flexibility that is required to address the rich and complex nature of the information that is now available to users. Morville writes:

We classify to understand...In a formal taxonomy, a single root node sits atop the hierarchy. Properties flow from class to subclass through the principle of inheritance. Each object and category is assigned a single location within the taxonomy. We live at an address within a nested hierarchy of streets, cities, states, and countries. We exist as *Homo sapiens* within the taxa of domain, kingdom, phylum, subphylum, class, order, family, genus, and species.

Of course, the world doesn't always cooperate with this Platonic approach to classification. Fish with lungs. Mammals that lay eggs. Documents about multiple topics. Words with many meanings. Meanings with many words. Reality confounds mutually exclusive classifications, and so we find ourselves debating which existing category works best or defining new categories to allow a perfect fit...

We embrace faceted classification...using multiple fields or "facets" to describe the objects within our collections. First defined in the 1930s by Indian librarian S. R. Ranganathan, faceted classifications have flourished in digital domains, where objects can exist simultaneously in many locations ([5], pp. 127-128).

Faceted classification and clustering differ in the way they group result lists, and each may be more applicable than the other in a particular context; faceted classification is a better fit when complete, comprehensive metadata is available and when groups can be predefined, whereas clustering is more suitable when metadata is incomplete or lacking or when the heterogeneous nature of the searched collection will benefit from a more flexible and dynamic method of determining groupings. "Rather than stuffing content into mutually exclusive buckets," as Peter Morville puts it, "we apply structural and semantic metadata" ([5], p. 128), thus helping users arrive at an intuitive understanding of what the result list offers and assisting them in navigating through the list more successfully.

ACKNOWLEDGEMENTS

The author thanks Nancy Dushkin for her encouragement and Ricka S. Rak for her wise advice and continual support.

⁴ [OAIster](#) is a union catalog of digital resources. Metadata records in OAIster are harvested from multiple digital resources using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

REFERENCES

- [1] K. Markey, 'Twenty-five years of end-user searching, Part 2: Future research directions', *Journal of the American Society for Information Science and Technology*, 58 (2007) 1123-1130.
<http://portal.acm.org/citation.cfm?id=1254879>.
- [2] K. Markey, 'The online library catalog: Paradise lost and paradise regained?', *D-Lib Magazine*, 13, 1/2 (2007).
<http://www.dlib.org/dlib/january07/markey/01markey.html>.
- [3] J. Griffiths and P. Brophy, 'Student searching behavior and the Web: Use of academic resources and Google', *Library Trends*, 53 (2005) 539-554.
http://findarticles.com/p/articles/mi_m1387/is_4_53/ai_n14732768.
- [4] Centre for Information Behaviour and the Evaluation of Research (CIBER). *Information behaviour of the researcher of the future: executive summary*. A British Library and JISC study. (2008).
<http://www.ucl.ac.uk/slais/research/ciber/downloads/ggexecutive.pdf>.
- [5] Peter Morville, *Ambient Findability*. (Sebastopol, CA, 2005).
- [6] K. Hagedorn, S. Chapman, and D. Newman, 'Enhancing search and browse using automated clustering of subject metadata', *D-Lib Magazine*, 13, 7/8 (2007).
<http://www.dlib.org/dlib/july07/hagedorn/07hagedorn.html>.
- [7] D. Rosen, *Ex Libris PRIMO Usability Evaluations*, University of Minnesota Usability Services (Minneapolis, 2006).
- [8] D. Rosen, *Ex Libris Primo Round 2 Usability Evaluations Summary Report*, University of Minnesota Usability Services (Minneapolis, 2007).
- [9] Mika Käki, 'Findex: Search result categories help users when document rankings fail', *Proceedings of ACM SIGCHI conference on human factors in computing systems* (Portland, 2005) 131-140.
<http://portal.acm.org/citation.cfm?doid=1054991>.
- [10] M. A. Hearst, 'Clustering versus faceted categories for information exploration', *Communications of the ACM*, 49, 4 (2006) 59-61.
<http://portal.acm.org/citation.cfm?doid=1121949.1121983>.
- [11] Tamar Sadeh, 'The challenge of metasearching', *New Library World*, 105, 1198/1199 (2004) 104-112.
<http://dandini.emeraldinsight.com/Insight/viewContentItem.do?contentType=Article&contentId=1509164>.

WEB SITES REFERRED TO IN THE TEXT

Amazon: <http://www.amazon.com/>
AquaBrowser Library (Medialab Solutions): <http://www.medialab.nl/>
Boston College libraries: <http://www.bc.edu/libraries/>
Boston College libraries Super Sleuth: <http://www.bc.edu/supersleuth>
eBay: <http://www.ebay.com/>
Elsevier ScienceDirect: <http://www.sciencedirect.com>
Elsevier Scirus: <http://www.scirus.com/>
Endeca Technologies: <http://www.endeca.com/>
Ex Libris Group: <http://www.exlibrisgroup.com/>
Ex Libris Primo: <http://www.exlibrisgroup.com/category/PrimoOverview>
Google: <http://www.google.com>
Google Scholar: <http://scholar.google.com>
Harvard University HOLLIS Catalog: <http://lms01.harvard.edu/F>
Library of Congress Authorities: <http://authorities.loc.gov/>
Mercado: <http://www.mercado.com/>
North Carolina State University catalog: <http://www.lib.ncsu.edu/catalog/>
OAster: <http://www.oaister.org/>
University of Iowa libraries Smart Search: <http://smartsearch.uiowa.edu>
University of Michigan MBooks: <http://www.lib.umich.edu/mdp/>
Vivísimo: <http://vivisimo.com/>
Vivísimo Clusty: <http://clusty.com/>
Vivísimo Velocity Enterprise Search: <http://vivisimo.com/html/products>
WorldCat: <http://worldcat.org/>