A User Study of Multi-collection Text Retrieval

Matt Stoeffler
Digital Library Production Service
University of Michigan
Ann Arbor, MI  48103
+1 734 615-2484

James Reed
Scholarly Publishing Office
University of Michigan
Ann Arbor, MI  48103
+1 734 615-7617

ABSTRACT
An ideal digital library should provide powerful text analysis tools with the ability to search multiple, varied collections simultaneously. This report describes the results of a usability study for the University of Michigan Digital Library (UMDL) text retrieval system. The primary goals of the study were to learn how people search across multiple text collections, and to identify problems in the UMDL search interface. In the study, eight participants were asked to complete four tasks that involved the selection of single or multiple collections for searching. The study revealed that people encountered severe difficulties in the following two areas. Firstly, people found it difficult to select groups of collections to search. Secondly, people had trouble specifying complex search parameters. Since digital libraries continue to offer more texts, greater text variety, and additional search options, search interfaces should provide additional support for these tasks.

Keywords
Digital Libraries, user testing, speak-aloud protocol

INTRODUCTION
The UMDL Texts site presents users with more than 25 “collections” of monographic texts encompassing a variety of subject areas and periods, and encoded in SGML at differing levels of granularity. The differing levels of data capture result in variations in search options, details available in results displays, and full text viewing. From the UMDL Texts site, users can select any combination of text collections to search simultaneously, or they can opt to navigate to the home site of any individual collection and search only within the items in that collection. In either circumstance, users may also, as they view search results or the full text, select items to save in their “book bag.” Book bag items can also be searched collectively at any time.

While many of the search options, results and text displays have been available for some time in DLPS retrieval systems, much of the cross collection functionality was new.

Our goal was to learn how users perform search and retrieval tasks in a multi-collection environment, and to identify potential problems in the UMDL Texts interface. Some general questions guided our inquiry: How easy or hard is it to compose a search against multiple collections using the current search interface? How would users interpret multiple collection results display formats? How clearly would they identify the needed information?

BACKGROUND
The UMDL Texts site presents users with more than 25 “collections” of monographic texts encompassing a variety of subject areas and periods, and encoded in SGML at differing levels of granularity. The differing levels of data capture result in variations in search options, details available in results displays, and full text viewing. From the UMDL Texts site, users can select any combination of text collections to search simultaneously, or they can opt to navigate to the home site of any individual collection and search only within the items in that collection. In either circumstance, users may also, as they view search results or the full text, select items to save in their “book bag.” Book bag items can also be searched collectively at any time.

METHODS
Eight users (four men and four women) participated in the testing. Users were all recruited via invitations sent from public service librarians from a variety of subject specializations. Among the users were 5 students ranging in status from undergraduate to Ph.D. Three of the users were librarians and all but one worked in a library.

All users were positioned in front of a Macintosh G3 Powerbook with 1024 X 768 screen resolution. Users’ screen actions were recorded on video tape which was later reviewed, creating a log of observable errors. Each user was asked to complete 4 tasks. The tasks were
printed out individually on standard sized paper and handed to the user at the beginning of each task. A speak aloud protocol was also employed, encouraging users, while performing their tasks, to verbally describe what they thought was happening.

The tasks were as follows:

Task #1: Verify some information about Oscar Wilde’s *The Importance of Being Earnest*:

Who is the publisher?
How many parts are cast?
What is the setting for the first scene in the first act?

Task #2: You are interested in exploring 19th century views of the Kansas-Nebraska Act of 1854. Can you find 3 documents that might be relevant to your research? Add the relevant documents to your book bag. (Hint: try looking in the Making of America collection.)

Task #3: You are interested in references to John and/or Robert Kennedy in 20th century American poetry. Find poems by American authors that contain references to one or the other or both.

Task #4: You are studying William Carlos Williams' use of natural imagery. Find variations of the term "leaf" in his works. (2 Hints: look in the Database of 20th Century American Poetry and use your book bag.)

Each user participated in a post-test debriefing interview, in which the procedure itself was discussed, as well as selected moments from any of the tasks themselves. Finally, each user completed a brief profile questionnaire to indicate their job status, as well as their relative level of web experience and competency.

RESULTS AND DISCUSSION
All but one user completed all the tasks, although task #4 was problematic: none of the users followed an ideal path in completing this task. Through the course of the tasks, however, a number of problem areas were identified where user error rates were significant, or where users’ verbal comments identified a problem.

The cross collection search page
The cross collection “Simple search” consists of two main components: At the top is the most basic term searching form, in which a user can enter any word or phrase to match in a specified field; the second component, which is part of all cross collection search dialogs, is an interface for specifying collections to include in the search. The form provides users with two means of picking collections. They can either manually check or uncheck boxes next to collection names, or they can select values in categorical filters (e.g., subject/genre, period, language, gender). Selecting values among the filters dynamically alters the checkboxes next to collections to reflect their correspondence with the value so selected. Users can also uncheck boxes manually after selecting values from the filters. By default, all available collections are included, as indicated by their associated checkboxes being checked, so that if they choose not to make any changes to the form, all searches will be conducted against all collections. An additional consequence of selecting or unselecting collections for the search is that the available search areas of the text in the upper form (in a pull down box next to the input box) change. This is because various collections have more data in common. When all collections are selected, the only area in common is “full text” which is the top-most level of information in the text and header.

Tasks two through four presented the user with a situation in which they might want to make restrictions to collections, and the form proved difficult for many of them. On tasks two and three there were 15 errors related to the collection selection interface, involving a third and half of the users respectively. All but one of these errors involved confusion about the relationship between the collection list and the collection filters: commonly a user selected a collection in the list, then selected values from the filters, which required them to again re-select collections in the list. In addition, half the users commented that sticky collection selections, where a collection selected in a previous search remained in a new search, were unexpected and frustrating.

The results display
The results display for cross collection searches is divided vertically between a results summary panel on the left, in which results numbers for each selected collection are shown, and the results display panel taking up the majority (right portion) of the page, in which results are indicated for each item in the collection in which a match occurred. Occurrence numbers on the left panel are linked and serve as a means for navigating between results for each separate collection included in the search. Results in the right panel can display in three primary ways, depending on the search type and richness of the document encoding:

1. Hierarchical headers with keyword in context (KWIC) information, reflect the results of a simple term search against a richly encoded text;
2. Hierarchical headers without KWIC information reflect the results of a Boolean type search against a richly encoded text;
3. Item titles with summary hit information in parentheses (i.e., nn hits in nn of nn pages ), reflect results details when the matches occur in a collection that is not as richly encoded.

The variance in results display had some minor negative impact on users’ ability to quickly interpret the information in the display and make subsequent decisions.
In tasks 2 and 4 there were 9 errors involving reading and interpreting results from Boolean searches, where only section headings are shown. In these cases, users were confused by the occasionally too generic nature of linked header content for the matching region (e.g., “section” or “paragraph”), and so were not clear which link to click. At the same time, they frequently noted their confusion as to why results from Boolean type searches did not display KWIC information. In the case of title only displays, where within-item results details are one click down from the “hits in document” link, some users were expecting to see KWIC information at the lower level and were frustrated when it was not found. Another problem area here was the “table of contents” links in results pages, which did not clearly indicate that bibliographic information was also available.

Another area of concern was the determination of results relevancy and the refine search function. 3 users commented on the difficulty of determining relevance in search results from less richly encoded collections, either because they did not see the summary hits information in parentheses, or because they expected to see or have available results in relevance order. In addition, users committed 5 errors in cases where they clicked “refine” expecting to search within their current found set, when in fact the link returns to an advanced search form where they can add or remove terms from their current search. It was common for users to react to a found set showing 40 or more items by immediately attempting to restrict the found set to a more manageable size.

**Simple or Boolean search forms**
The overwhelming factor on all search forms was the lack of ability to include multiple fields in “Boolean” searches, and especially the ability to include author and title information. 5 of 8 users commented that they expected to include author in advanced searches. Another lesser problem with search forms included confusing labeling of areas of text to which users could direct their queries. One label in particular was “works”, which actually corresponds to a composite of all lower level structures in the text. 3 of 8 users commented that they thought the “works” option seemed ambiguous.

**Miscellaneous other problem areas**
The last most significant discovery, coming from users’ verbal comments, was that all but one user did not use the citation search form, either because they never saw the citation search link in the secondary navigation bar on advanced search pages, or that the label “citation” did not clearly indicate what the search was for. Both causes are probable and not necessarily exclusive. The expressed desire for author and title restrictions to all other searches indicates, it seems, a different model of searching options than the current one in which bibliographic information is only searched exclusively of all other categories.

**RECOMMENDATIONS**
Results of the study for the UMDL Texts interface indicate that the following recommended changes be considered to improve ease of use and / or effectiveness of the resource:

The collection selection interface should be redesigned to reduce the number of simultaneous form interactions. Quite probably, it is necessary to remove the selection of collections to another page entirely. As a supplement, grouping of collections from which to make selections could be re-conceptualized, so as to streamline the process of selection for the user.

Revise search forms to include restrictions for bibliographic information, such as author and title.

Look for ways to improve user perception of search information in search results pages. The simplest approach: Improve instruction text to communicate more clearly the information available on search pages containing details for collections encoded at any level. The more involved approach: Revise the overall model of results display to enforce greater consistency across all encoding levels, listing consistent item level information at the first depth level of results. This would ensure easier comparison of results across all searches, but would require significant exploration as to how richer levels of information should be made available in list format while maintaining consistency of function.

Reevaluate the “refine” search functionality to clearly differentiate the process of editing original search terms against all selected collections form the process of submitting a search against the current found set. Perhaps it is a good idea to develop the capability to perform both functions.

**SUMMARY**
This report describes the results of a user study of a multi-collection search interface for UMDL texts. The study indicates that users faced severe difficulties in two main areas:

- People found the interface for selecting collections in the initial steps of a cross collection search difficult to operate correctly, and
- People were consistently frustrated with the selection of search types, and specifically by the inability to restrict any term search further with author and/or title words.
Other lesser problems included some difficulty interpreting specific label text in form menus in which users selected target areas of a text to search, difficulties determining relevancy in some search results displays, and the lack of use of the “Citation Search” form.

Based on study findings, recommended changes to be considered include the following:

Simplify the search process. The collection selection interface in UMDL Texts should be redesigned to reduce the number of simultaneous form interactions.

Grouping of collections from which to make selections could be re-conceptualized to create broader, generalizable categories that users could more easily digest.

UMDL Texts search forms should be redesigned to include restrictions for bibliographic information, such as author and title.