

Group 2  
(names withheld for privacy)

Assignment 2  
Subject Analysis

Part A: Database Design
-------------------------

**1. A user guide, as described in #5 above.**

This database is an online collection of articles from various journals about information retrieval and other related topics.

The primary actual user group of this database is students enrolled in any section of San Jose State University's Libr 202: Information Retrieval course, with an expected user group of other SJSU MLIS students and faculty researching information retrieval topics. This database allows users to search for articles by standard features such as title and author, as well as by topic via different types of index terms. Because the user group is mainly comprised of library science students, we can assume that users are familiar with the DBTextWorks software.

How to search this database:

This database permits searching in a variety of ways.

To search by **Author**, enter the last name first. The search is not case sensitive. You may search for one or multiple authors.

**Title** may be entered in full (leaving off any introductory articles, such as “the”), or users may conduct a partial title or keyword search within the title.

**Journal** Title can be searched in the same way: full title, partial title, or keyword within the title.

**Volume** allows a user to find articles from a particular issue. Volume is abbreviated “v,” Number is abbreviated “no” and Arabic numbering is used. Users must enter the appropriate abbreviations, no spaces: “no5.”

**Pages** allows you to search for an article if you know which pages of the journal it was printed on. It does not allow you to search by length of pages per article. Page numbers are preceded by the abbreviation “pp” (no punctuation) and then the pagination. Users may enter only the starting page number if that is all that is known.

Users may search by **Date**, using the month (spelled out in its entirety, no abbreviations) and then the year, e.g., “January 1986.” Month or year alone may also be entered.

Each record contains an **Abstract**, or short summary, of the article. Users may enter keywords or combinations of keywords to search within the abstracts.

The **Preco** and **Postco** fields allow a user to search by concepts and index terms. Both sets of terms use plural instead of singular and -ing endings when appropriate. The Preco field (short for “pre-coordinate indexing”) contains strings of terms, tapering down to very specific concepts. Terms are narrowed by dashes (two hyphens: e.g., “topic—subdivision”). Users must enter the entire string. This is useful for searching for very specific topics. The Postco field (short for “post-coordinate indexing”) also allows users to search by index terms. Unlike the Preco field, the Postco terms are much broader and do not string or narrow down. This allows users more flexibility and the ability to return a broad range of subtopics on a given subject.

## 2. The data structure, inserted as a file from DBTextWorks into your Word document.

### Textbase Structure

Textbase: C:\Documents and Settings\Ryan\Desktop\assignment 2\assignment  
2

Created: 5/4/2006 1:20:17 AM

Modified: 5/10/2006 2:07:19 AM

### Field Summary:

1. ID number: Automatic Number(next avail=17, increm=1), Term
2. Author: Text, Term & Word  
Validation: required
3. Title: Text, Term & Word  
Validation: required
4. Journal: Text, Term & Word
5. Volume: Text, Term & Word
6. Pages: Text, Term & Word
7. Date: Text, Term & Word  
Validation: required
8. Abstract: Text, Term & Word  
Validation: required
9. Postco: Text, Term  
Thesaurus: C:\Documents and Settings\Ryan\Desktop\assignment  
2\thesaurus  
Validation: required, valid-list(thesaurus)
10. Preco: Text, Term  
Thesaurus: C:\Documents and Settings\Ryan\Desktop\assignment  
2\thesaurus  
Validation: required, valid-list(thesaurus)

Log file enabled, showing 'ID number'

Leading articles: a an the

Stop words: a an and by for from in of the to

### Textbase Defaults:

Default indexing mode: SHARED IMMEDIATE

Default sort order: <none>

### Textbase passwords:

Master password = ''

0 Access passwords:

No Silent password

3. The validation lists for your preco and postco fields, inserted as files from DBTextWorks into your Word document

Thesaurus('thesaurus') for field 'Postco', textbase 'assignment 2',  
5/10/2006 3:13:42 AM:

academic libraries  
artificial intelligence  
bibliographic database searching  
bibliographic instruction  
cognitive model  
cognitive process  
colinked descriptors USE thesauri  
college and university libraries USE academic libraries  
college libraries USE academic libraries  
computer-user interaction USE human computer interaction  
controlled vocabulary USE thesauri  
design  
evaluation  
full-text  
human behavior  
human computer interaction  
indexing  
information  
information processing  
information retrieval  
information seeking  
interfaces  
IR USE information retrieval  
keyword USE natural language  
knowledge organization  
library catalogs USE online library catalogs  
mental model  
methods  
natural language  
online  
online bibliographic searching USE bibliographic database searching  
online catalogs USE online library catalogs  
online library catalogs  
online searching USE bibliographic database searching  
relevance  
search strategies  
subject access  
subject cataloging  
systems  
thesauri  
thesaurus USE thesauri  
university libraries USE academic libraries  
users

Thesaurus('thesaurus') for field 'Preco', textbase 'assignment 2',  
5/10/2006 3:14:30 AM:

academic libraries  
artificial intelligence  
artificial intelligence -- systems -- expert  
bibliographic database searching  
bibliographic database searching -- methods  
bibliographic instruction  
bibliographic instruction -- academic libraries -- evaluation  
cognitive model

cognitive process  
cognitive process -- human  
colinked descriptors USE thesauri  
college and university libraries USE academic libraries  
college libraries USE academic libraries  
computer-user interaction USE human computer interaction  
controlled vocabulary USE thesauri  
design  
evaluation  
full-text  
human behavior  
human computer interaction  
indexing  
indexing -- evaluation  
indexing process -- cognitive model  
information  
information processing  
information processing -- psychology  
information retrieval  
information retrieval -- evaluation  
information retrieval -- full-text relevance  
information retrieval -- relevance  
information retrieval -- systems  
information retrieval -- systems -- design  
information retrieval -- systems -- online evaluation  
information seeking  
interfaces  
interfaces -- user-friendly  
IR USE information retrieval  
keyword USE natural language  
keyword searching USE natural language -- input  
knowledge organization  
knowledge organization -- human  
library catalogs USE online library catalogs  
mental model  
methods  
natural language  
natural language -- input  
online  
online bibliographic searching USE bibliographic database searching  
online catalogs USE online library catalogs  
online library catalogs  
online library catalogs -- design  
online library catalogs -- design -- evaluation  
online library catalogs -- subject access  
online searching USE bibliographic database searching  
relevance  
search strategies  
subject access  
subject access -- evaluation  
subject cataloging  
systems  
thesauri  
thesaurus USE thesauri  
university libraries USE academic libraries  
users

Term index for field 'Postco', textbase 'assignment 2', 5/11/2006  
1:40:26 PM:

1           academic libraries

1 artificial intelligence  
2 bibliographic database searching  
1 bibliographic instruction  
1 cognitive model  
3 cognitive process  
5 design  
4 evaluation  
1 full text  
1 human behavior  
2 human computer interaction  
2 indexing  
1 information  
1 information processing  
10 information retrieval  
1 information seeking  
4 interfaces  
1 knowledge organization  
1 mental model  
2 natural language  
1 online  
2 online library catalogs  
2 relevance  
2 search strategies  
2 subject access  
1 subject cataloging  
7 systems  
1 thesauri  
1 users

Total number of keys: 29

Term index for field 'Preco', textbase 'assignment 2', 5/11/2006 1:40:56 PM:

1 artificial intelligence systems expert  
2 bibliographic database searching  
1 bibliographic database searching methods  
1 bibliographic instruction academic libraries evaluation  
3 cognitive process human  
2 human computer interaction  
1 indexing evaluation  
1 indexing process cognitive model  
1 information processing psychology  
3 information retrieval  
1 information retrieval evaluation  
1 information retrieval full text relevance  
1 information retrieval relevance  
2 information retrieval systems  
3 information retrieval systems design  
1 information retrieval systems online evaluation  
3 interfaces user friendly  
1 knowledge organization human  
1 mental model  
1 natural language input  
1 online library catalogs design  
1 online library catalogs design evaluation  
1 online library catalogs subject access  
2 search strategies  
1 subject access evaluation  
1 subject cataloging  
1 thesauri

Total number of keys: 27

#### 4. A printout of your records, inserted as a file from DBTextWorks into your Word document.

ID number 1

Author Farrow, John F.

Title A Cognitive Process Model of Document Indexing

Journal Journal of Documentation

Volume v47, no2

Pages pp149-166

Date June 1991

Abstract Classification, indexing and abstracting can all be regarded as summarisations of the content of a document. A model of text comprehension by indexers (including classifiers and abstractors) is presented, based on task descriptions which indicate that the comprehension of text for indexing differs from normal fluent reading in respect of: operational time constraints, which lead to text being scanned rapidly for perceptual cues to aid gist comprehension; comprehension being task oriented rather than learning oriented, and being followed immediately by the production of an abstract, index, or classification; and the automaticity of processing of text by experienced indexers working within a restricted range of text types. The evidence for the interplay of perceptual and conceptual processing of text under conditions of rapid scanning is reviewed. The allocation of mental resources to text processing is discussed, and a cognitive process model of abstracting, indexing and classification is described.

Postco cognitive model  
indexing

Preco indexing process -- cognitive model

ID number 2

Author Huston, Mary M.

Title Windows into the Search Process: an Inquiry  
into Dimensions of Online Information Retrieval

Journal Online Review

Volume v15, no3/4

Pages pp227-243

Date June-August 1991

Abstract From diverse users' points of view, contextual frameworks are elaborated for the nature of the information technology, the information universe, and the information search. Within these conceptual parameters, established theories on search strategy are reviewed and cognitive models of information-seeking are highlighted. Future directions for research on users' search processes are discussed in terms of the role for online retrieval in the future information environment.

Postco information retrieval  
search strategies  
online

Preco information retrieval

bibliographic database searching  
search strategies  
ID number 3  
Author Ingwersen, Peter  
Title Cognitive Perspectives of Information  
Retrieval Interaction: Elements of a  
Cognitive IR Theory  
Journal Journal of Documentation  
Volume v52, no1  
Pages pp3-50  
Date September 1996

Abstract The objective of the paper is to amalgamate theories of text retrieval from various research traditions into a cognitive theory for information retrieval interaction. Set in a cognitive framework, the paper outlines the concept of polyrepresentation applied to both the user's cognitive space and the information space of IR systems. The concept seeks to represent the current user's information need, problem state, and domain work task or interest in a structure of causality. Further, it implies that we should apply different methods of representation and a variety of IR techniques of different cognitive and functional origin simultaneously to each semantic full-text entity in the information space. The cognitive differences imply that by applying cognitive overlaps of information objects, originating from different interpretations of such objects through time and by type, the degree of uncertainty inherent in IR is decreased. Polyrepresentation and the use of cognitive overlaps are associated with, but not identical to, data fusion in IR. By explicitly incorporating all the cognitive structures participating in the interactive communication processes during IR, the cognitive theory provides a comprehensive view of these processes. It encompasses the ad hoc theories of text retrieval and IR techniques hitherto developed in mainstream retrieval research. It has elements in common with van Rijsbergen and Lalmas' logical uncertainty theory [1] and may be regarded as compatible with that conception of IR. Epistemologically speaking, the theory views IR interaction as processes of cognition, potentially occurring in all the information processing components of IR, that may be applied, in particular, to the user in a situational context. The theory draws upon basic empirical results from information seeking investigations in the operational online environment, and from mainstream IR research on partial matching techniques and relevance feedback.

Postco cognitive process  
information retrieval  
interfaces

Preco cognitive process -- human  
information retrieval  
search strategies

ID number 4

Author Najarian, Suzanne E.

Title Organizational Factors in Human Memory:  
Implications for Library Organizations and  
Access Systems

Journal The Library Quarterly

Volume v51, no3

Pages pp269-291

Date 1981

Abstract Psychological studies on memory about human categorizing processes and the organizing principles and limitations of human memory. Particular attention is given to evidence for a mode 1 which represents the organization of knowledge in memory in terms of a hierarchical type of structure. The experimental findings suggest several considerations for the design of library systems of organization and access that would take into account characteristics of the conceptual organization of knowledge. Such systems are likely to be particularly effective in aiding the user in his search for information since they would (1) employ organizational schemes that are familiar to the individual, (2) permit a strategy for the exploration of a subject area similar to the type of search procedure which seems to facilitate the retrieval of items from memory, and (3) take into consideration the apparent limits on the amount of information that the individual can successfully attend to at one time.

Postco cognitive process  
design  
information  
systems  
knowledge organization

Preco cognitive process -- human  
knowledge organization -- human  
information retrieval -- systems -- design

ID number 5

Author Simon, Herbert A.

Title Information-Processing Models of Cognition  
Journal Journal of the American Society for  
Information Science

Volume no5

Pages pp364-377

Date September 1981

Abstract This article reviews recent progress in modeling human cognitive processes. Particular attention is paid to the use of computer programming languages as a formalism for modeling, and to computer simulation of the behavior of the systems modeled. Theories of human cognitive processes can be attempted at several levels: at the level of neural processes, at the level of elementary

information processes (e.g., retrieval from memory, scanning down lists in memory, comparing simple symbols, etc.), or at the level of higher mental processes (e.g., problem solving, concept attainment). This article will not deal at all with neural models; it focuses mainly upon higher mental processes, but not without some attention to modeling the elementary processes and especially to the relationships between elementary and complex processes.

Postco cognitive process  
human behavior  
information processing  
Preco information processing -- psychology  
cognitive process -- human

ID number 6

Author Ury, Connie Jo  
Johnson, Carolyn V.  
Meldrem, Joyce A.

Title Teaching a Heuristic Approach to Information  
Retrieval

Journal Research Strategies

Volume v15, no1

Pages pp39-47

Date Winter 1997

Abstract To become life-long learners, students must acquire information retrieval skills for future as well as current information needs. This article describes how the Library Use Instruction Program at Northwest Missouri State University incorporates a heuristic model in which students continually evaluate and refine their information seeking practices while progressing through all levels of courses in diverse disciplines. Collegial partnerships with departmental faculty and ongoing instructional assessment are essential to the success of the program.

Postco bibliographic instruction  
academic libraries  
evaluation  
information retrieval

Preco bibliographic instruction -- academic  
libraries -- evaluation  
information retrieval -- evaluation

ID number 7

Author Bates, Marcia J.

Title Indexing and Access for Digital Librarians and  
the Internet: Human, Database, and Domain  
Factor

Journal Journal of the American Society for  
Information Science

Volume v49, no13

Pages pp1185-1205

Date November 1998

Abstract Discussion in the research community and among the general public regarding content indexing (especially subject indexing) and access to digital resources, especially on the Internet, has underutilized research

on a variety of factors that are important in the design of such access mechanisms. Some of these factors and issues are reviewed and implications drawn for information system design in the era of electronic access. Specifically the following are discussed: Human factors: Subject searching vs. indexing, multiple terms of access, folk classification, basic-level terms, and folk access; Database factors: Bradford's Law, vocabulary scalability, the Resnikoff-Dolby 30:1 Rule; Domain factors: Role of domain in indexing.

Postco indexing  
    online library catalogs  
    subject access  
    design  
    evaluation  
Preco online library catalogs -- design  
    indexing -- evaluation  
    online library catalogs -- subject access

ID number 8

Author Bates, Marcia J.

Title Subject Access in Online Catalogs: A Design Model

Journal Journal of the American Society for Information Science

Volume v37, no6

Pages pp357-375

Date November 1986

Abstract       A model based on strikingly different philosophical assumptions from those currently popular is proposed for the design of online subject catalog access. Three design principles are presented and discussed: uncertainty (subject indexing is indeterminate and probabilistic beyond a certain point), variety (by Ashby's law of requisite variety, variety of searcher query must equal variety of document indexing), and complexity (the search process, particularly during the entry and orientation phases, is subtler and more complex, on several grounds, than current models assume). Design features presented are an access phase, including entry and orientation, a hunting phase, and a selection phase. An end-user thesaurus and a front-end system mind are presented as examples of online catalog system components to improve searcher success during entry and orientation.

The proposed model is "wrapped around" existing Library of Congress subject-heading indexing in such a way as to enhance access greatly without requiring reindexing. It is argued that both for cost reasons and in principle this is a superior approach to other design philosophies.

Postco subject cataloging  
    design

online library catalogs  
subject access  
evaluation  
Preco online library catalogs -- design -- evaluation  
subject cataloging  
subject access -- evaluation  
ID number 9  
Author Bates, Marcia J.  
Title The Design of Browsing and Berrypicking  
Techniques for the Online Search Interface  
Journal Online Review  
Volume v13, no5  
Pages pp407-424  
Date 1989  
Abstract First, a new model of searching in online and other information systems, called "berrypicking", is discussed. This model, it is argued, is much closer to the real behavior of information searchers than the traditional model of information retrieval is, and, consequently, will guide our thinking better in the design of effective interfaces. Second, the research literature of manual information seeking behavior is drawn on for suggestions of capabilities that users might like to have in online systems. Third, based on the new model and research on information seeking, suggestions are made for how new search capabilities could be incorporated into the design of search interfaces. Particular attention is given to the nature and types of browsing that can be facilitated.

Postco information retrieval  
search strategies  
systems  
interfaces  
bibliographic database searching  
Preco bibliographic database searching -- methods  
information retrieval -- systems  
interfaces -- user-friendly  
ID number 10  
Author Borgman, Christine L.  
Title The User's Mental Model of an Information Retrieval System: An Experiment on a Prototype Online Catalog  
Journal International Journal of Man-Machine Studies  
Volume v24  
Pages pp47-64  
Date January 1986  
Abstract An empirical study was performed to train naive subjects in the use of a prototype Boolean logic-based information retrieval system on a database of bibliographic records. The research was based on the mental models theory which proposes that people can be trained to develop a "mental model" or a qualitative simulation of a system which will aid in generating methods for interacting with the system, debugging errors, and

keeping track of one's place in the system. It follows that conceptual training based on a system model will be superior to procedural training based on the mechanics of the system. We performed a laboratory experiment with two training conditions (model and procedural), and with each condition split by sex. Forty-three subjects participated in the experiment, but only 32 were able to reach the minimum competency level required to complete the experiment. The data analysis incorporated time-stamped monitoring data, personal characteristics variables, affective variables, and interview data in which subjects described how they thought the system worked (an articulation of the model). As predicted, the model-based training had no effect on the ability to perform simple, procedural tasks, but subjects trained with a model performed better on complex tasks that required extrapolation from the basic operations of the system. A stochastic process analysis of search-state transitions reinforced this conclusion. Subjects had difficulty articulating a model of the system, and we found no differences in articulation by condition. The high number of subjects (26%) who were unable to pass the benchmark test indicates that the retrieval tasks were inherently difficult. More interestingly, those who dropped out were significantly more likely to be humanities or social science majors than science or engineering majors, suggesting important individual differences and equity issues. The sex-related differences were slight, although significant, and suggest future research questions.

Postco mental model  
information retrieval  
human computer interaction  
systems  
design  
Preco information retrieval -- systems -- design  
human computer interaction  
mental model

ID number 11

Author Gauch, Susan

Title Intelligent Information Retrieval: An Introduction

Journal Journal of the American Society for Information Science

Volume v43, no2

Pages pp164-174

Date March 1992

Abstract Researchers are exploring the application of artificial intelligence techniques to information retrieval with the goal of providing intelligent access to online information.

This article surveys several such systems to

show what is possible in the lab today, and what may be possible in the library or office of tomorrow. Systems incorporating user modeling, natural language understanding, and expert systems technology are presented.

Postco artificial intelligence  
natural language  
information retrieval  
human computer interaction  
systems

Preco human computer interaction  
information retrieval  
artificial intelligence -- systems -- expert  
natural language -- input

ID number 12

Author Harman, Donna

Title User-Friendly Systems Instead of User-Friendly Front-Ends

Journal Journal of the American Society of Information Science

Volume v43, no2

Pages pp164-174

Date March 1992

Abstract Most commercial online retrieval systems are not designed to service end users and, therefore, have often built "front-ends" to their systems specifically to serve the end-user market. These front-ends have not been well accepted, mostly because the underlying systems are still difficult for end users to use successfully in searching. New techniques, based on statistical methods, that allow natural language input and return lists of records in order of likely relevance, have long been available from research laboratories. This article presents four prototype implementations of these statistical retrieval systems that demonstrate their potential as powerful and easily used retrieval systems able to service all users.

Postco information retrieval  
systems  
interfaces  
natural language

Preco information retrieval -- systems  
interfaces -- user-friendly

ID number 13

Author Harter, Stephen P.  
Cheng, Yung-Rang

Title Colinked Descriptors: Improving Vocabulary Selection for End-User Searching

Journal Journal of the American Society of Information Science

Volume v47, no4

Pages pp311-325

Date April 1996

Abstract This article introduces a new concept and technique for information retrieval called colinked descriptors. Borrowed from an analogous idea in bibliometrics-cocited references colinked descriptors provide a

theory and method for identifying search terms that, by hypothesis, will be superior to those entered initially by a searcher. The theory suggests a means of moving automatically from two or more initial search terms, to other terms that should be superior in retrieval performance to the two original terms. A research project designed to test this colinked descriptor hypothesis is reported. The results suggest that the approach is effective, although methodological problems in testing the idea are reported. Algorithms to generate Co-linked descriptors can be incorporated easily into system interfaces, front-end or pre-search systems, or help software, in any database that employs a thesaurus. The potential use of colinked descriptors is a strong argument for building richer and more complex thesauri that reflect as many legitimate links among descriptors as possible.

Postco thesauri  
information retrieval  
bibliographic database searching  
relevance  
Preco information retrieval -- relevance  
bibliographic database searching  
thesauri

ID number 14

Author Marchionini, Gary

Title Interfaces for End-User Information Seeking  
Journal Journal of the American Society for  
Information Science

Volume v43, no2

Pages pp156-163

Date March 1992

Abstract Essential features of interfaces to support end-user information seeking are discussed and illustrated. Examples of interfaces to support the following basic information-seeking functions are presented: problem definition, source selection, problem articulation, examination of results, and information extraction. It is argued that present interfaces focus on problem articulation and examination of results functions, and research and development are needed to support the problem definition and information extraction functions. General recommendations for research on interfaces to support end-user information seeking include: attention to multimedia information sources, development of interfaces that integrate information-seeking functions, support for collaborative information seeking, use of multiple input/output devices in parallel, integration of advanced information retrieval techniques in systems for end users, and development of adaptable interfaces to meet individual difference and multicultural needs.

Postco users

interfaces  
information seeking  
information retrieval  
systems  
design  
Preco interfaces -- user-friendly  
information retrieval -- systems -- design  
ID number 15  
Author Maron, M.E.; Blair, David C.  
Title An Evaluation of Retrieval Effectiveness for a  
Full-Text Document-Retrieval System  
Journal Communications of the ACM  
Volume v28, no3  
Pages pp289-299  
Date March 1985  
Abstract An evaluation of a large, operational full-  
text document-retrieval system (containing  
roughly 350,000 pages of text) shows the  
system to be reentering less than 20 percent  
of the documents relevant to a particular  
search. The findings are discussed in terms  
of the theory- and practice of full-text  
document retrieval.  
Postco information retrieval  
full-text  
relevance  
evaluation  
systems  
Preco information retrieval -- full-text relevance  
information retrieval -- systems -- online  
evaluation

**1. Write up a description of your criteria and the rationale for them. Be creative; go for major concepts that are grounded in your readings. Be sure to attribute concepts taken from your readings and list those readings in your list of references at the end.**

There are many criteria used to evaluate information retrieval systems. Below are definitions of information retrieval system criteria used to evaluate this particular textbase:

**Coverage:** the proportion of the total potentially useful literature that has been analyzed. For this sample textbase evaluation, the coverage is 100%.

**Recall:** a measure of completeness of retrieval. The equation for recall is:

$$\frac{\text{Relevant items returned}}{(\text{relevant items returned} + \text{relevant items not returned})}$$

**Precision:** a measure of accuracy of retrieval. The equation for precision is:

$$\frac{\text{Relevant items retrieved}}{(\text{relevant items retrieved} + \text{non-relevant items retrieved})}$$

Recall and precision are affected by indexing exhaustivity and term specificity. Higher level of exhaustivity of indexing might ensure high recall but decrease the level of precision. The system will retrieve a large number of non-relevant documents. High level of term specificity tends to ensure high precision but decrease the level of recall, creating an inverse relationship.

**Fallout:** a measure of what proportion of non-relevant items has been retrieved in a given search. It can measure the exhaustivity of indexing and term specificity.

In addition to the criteria above, there are some user-oriented evaluation measures:

**Relevance:** The equations for recall and precision all contain the concept “relevance,” therefore it is necessary to judge how many records are relevant in the database to the request and which retrieved record is relevant to the request before calculating recall and precision. For this assignment, we use the concept “on the same topic” to define relevance. The textbase evaluator set up the requests and judged whether the retrieved documents are on the same topic of the requests or not. Usefulness is not considered in this portion of the evaluation process.

**Usability and presentation:** usability means the value of the references retrieved and presentation means the form in which search results are presented to the user. Usability and presentation are evaluated via user feedback and commentary.

**2. Using the criteria that you have developed, evaluate subject access in your database. Compare and evaluate the 4 points of subject access (the two controlled vocabularies and the 2 natural language fields: title and abstract).**

*2a. Describe your testing. Show the searches that you did and their results.*

First, statistical retrieval tests were run on the database. Terms were selected as sample queries. For the statistical purpose of recall, precision and fallout, the ideal terms should be used in more than one record and used in different subject fields. The pool of request terms should include both general terms and specific terms. Four terms were selected:

Term A: “natural language,”

Term B: “interfaces”

Term C: “information retrieval”

Term D: “information retrieval system”

Then, relevance was determined for each term according to the above definition. The record numbers of each relevant article for each term were recorded in the first column in Table A.

A search was run in the textbase using each request term in four subject fields: title, abstract, preco and postco. The record number of retrieved articles for each term per field are recorded in the corresponding cells in Table A.

	Relevant Records	Title	Abstract	Preco	Postco
Term A	11, 12		11, 12		11, 12
Term B	9, 12, 14	14	9, 13, 14		3, 9, 12, 14
Term C	1-4, 6-15	2, 3, 6, 10, 11	3, 6, 9, 10, 11, 13, 14	2, 3, 11	2, 3, 6, 9, 10-15
Term D	4, 7, 9, 10, 11, 12, 15			9, 12	

Fourth, recall, precision and fallout were calculated for each term (Table B).

	Recall				Precision				Fallout			
	Title	Abs	Pre	Post	Ti	Abs	Pre	Post	Ti	Abs	Pre	Post
Term A		100		100		100		100				
Term B	33	66		100	100	66		100		8		8
Term C	36	50	21	71	100	100	100	100				
Term D			29				100					
Average	17	54	12.5	68	100	89	100	100				

In addition to statistical tests, feedback on the database was sought from external users. Criteria that user feedback sought to address included the following:

- Can the user identify the purpose of the database?
- Are the instructions easy for the user to understand?
- Can the user easily access the vocabulary standards?
- Are the vocabularies standards clear for the user?
- Is the database structured in a logical, clear format?
- Can the user easily identify the best fields to use to find information?
- Can the user find the information needed?

Two test users were consulted. Neither user was a library science student nor familiar with the field, although one test user was familiar with both creation and use of databases. The user guide (see Part A) was given to test users as well as ideas and prompts for searches, including terms, titles, and specific articles.

Replies from both test users indicated they were able to identify and understand the purpose of the database. One user found the introduction slightly misleading by saying this was a collection of online articles. The user expected to access the full text of the articles directly from the database. The user suggested stating it was a collection of abstracts of articles would be clearer. Both users found the instructions easy to understand, however, although both users found the vocabulary standards clear in the instructions, they both were still unclear about how to search the Postco and Preco fields. Including an example was suggested.

There was a disparity between test users as to whether the database was structured in a logical and clear format. One user found the order of the fields confusing and the labels unclear. The other user found the fields clear, but was confused as to the inclusion of some of the fields in the Query screen

The instructions were clear and easy to understand. One user did feel ID number needed to be included in the user guide for clarification of its role in the database. The other user understood the concept of the ID number, but inquired as to why it appeared in the Query form; the same for the "Pages" field.

The vocabulary standards are easily located in the instructions but, once at the query screen, there was no list to choose from. One user made multiple attempts at searching in title, abstract and postco fields for the words "full text." All searches were unsuccessful. The other user also experienced difficulty choosing terms, attempting examples like "information retrieval system" vs. "information retrieval systems."

The first user was not very clear on which fields to search first. This user attempted to locate articles by Buckland and retrieved no results. When the user finally

retrieved results from the database, the user was unsure how to interpret the results. The second user understood and had no problems with the Title, Author, or Abstract fields. This user could not understand how to run a search by pagination. He also had difficulty understanding the Preco and Postco fields.

The results regarding users fulfilling their information needs varied greatly. The first user was only able to successfully retrieve records using “indexing” to search the abstract field. All other searches performed in the title, journal, postco, and preco fields failed. The second user was successful in finding articles via the text fields, such as Title and Abstract, but had great difficulty, especially with the Preco field. This user searched the term “information retrieval” in the Preco field and got no results. Knowing they were strings but not knowing any additional terms, he then tried “information retrieval\*” (wildcard), which brought up 12 records.

*2b. Analyze the results of your searches. What did your tests show about the retrieval capabilities of the various fields? What are the problems and the advantages of each type of field? One approach is to consider the value added: Did each field contribute to retrieval commensurate with the cost of creating it? How do the various fields contribute to retrieval?*

### **Analysis of Table A**

Term A: There are two records (11, 12) in the textbase that are relevant to Term A (“natural language”). Term A is not present in the Title field in either record but it does appear in the Abstract and Postco fields in both records. Term A does appear in the Preco field, but as part of a string: “natural language – XXX,” so it is irretrievable without including the “XXX” in searching.

Term B: Three records (9, 12, 14) are relevant to term B (“interfaces”) in the database. When searching “interfaces” in the Title field, record 14 is retrieved. But if searching on the singular form “interface,” record 9 is retrieved. In the Abstract field, records 9, 13, and 14 were retrieved, but record 13 is not actually relevant to Term B. When searching the Preco field nothing was retrieved because term B is in the string “interfaces—user-friendly” in this field. In the Postco field, record 3 was retrieved in addition to the three relevant records because the indexer assigned the wrong subject term.

Term C: All the records except record 5 are relevant to term C (“information retrieval”). 5 records were retrieved by searching the Title field and 7 by searching the Abstract field. In the Preco field, only 3 records were retrieved because either the term was not assigned or because it was in a string as “information retrieval – XXX.”

Term D: A total of 7 records (4, 7, 9, 10, 11, 12, 15) are relevant to term D (“information retrieval systems”) in the database. No records were retrieved in the Title, Author or Postco fields because the term is not present in these fields. However, “information retrieval system” (not plural) retrieved record 10 when searching the Title and Abstract fields. Records 9 and 12 were retrieved when searching the Preco field. None of the additional relevant records were retrieved either because the indexer didn’t assign this

term (record 7) or because the term is in a string “information retrieval-system-XXX” (record 4, 10, 11, 15).

### **Analysis of Table B**

Recall: From Table B, the lowest recall is the Preco field. From the information in Table A, it can be determined that the reason Preco has the lowest recall is because the request terms are always in strings and this textbase software doesn't provide support for such terms, at least not in the design of this particular textbase. The Postco field had the highest recall because the indexing terms in this field were designed by indexer after reviewing the articles carefully and assigning more general terms. The 4 testing words selected leaned toward the general side, so they matched easily. The Title and Abstract fields are natural language subject fields. Because the authors of the articles usually try to express the main idea in these two areas, they contribute to higher recall. Generalization and singular vs. plural form of the request terms also affects the result.

Precision: Precision was found to be very high for all four terms and fields in this test. The 15 articles pre-selected by the system designer constitute a small sample. The request terms were selected by the tester who was familiar with the articles, and relevance was determined by the same tester. This accounts for the high precision rate.

Fallout: The reason of low fallout is the same as high precision.

As far as user-based testing, there are quite a variety of factors that affected the failure or success of the database usability. While testers agreed that the instructions for searching the database were clear, they were still unclear about the concepts behind some of the fields. This can be attributed to both a lack of clarity in the instructions and also a lack of familiarity on the test subjects' parts, as neither was familiar with the discipline of library science. A more focused test group made up of library science students and scholar could possibly give different results, and help to focus in on an improved description in the instructions.

The database Query screen itself posed a few problems for the test users. One found the order of the fields confusing, while the other questioned the inclusion of some of the fields. Both had questions regarding the ID number. To improve the database, the ID number should not be included in the Query screen, and only in the display of results if deemed necessary. The ID number is of little value to the end user and therefore does not need to appear in the results. Similarly, the “Pages” field, while necessary for citation or location once the article is retrieved, is mostly useless when searching. It is unlikely that a user will ever search for an article by page number. The pagination should be included in the record display, but it is not necessary to include it in the query screen.

Vocabulary caused a bit of concern. Consistency was lacking, and users could not access a list of standard vocabulary from which to choose terms. This was especially problematic in the term fields (Preco, Postco) where terms had to match exactly in order for retrieval. Both users had difficulty with the Preco strings. The entire string must be

entered (i.e., “information retrieval—systems—design”) in order to return the record. One user stumbled onto success by using the asterisk (\*) as a wildcard function, allowing the return of all strings beginning with “information retrieval.”

*2c. How would you improve the vocabulary, syntax, or other characteristics of the 4 fields? What would you do differently if you were going to continue to develop and maintain this database?*

Both statistical and user-based testing revealed many flaws in this database that could be improved upon.

Improvements in vocabulary control are key to improving this database. Vocabulary control is imperative within the database, with a better and more accessible validation list that both links to terms and is also accessible to users so that they may select appropriate terms from the list. The failure of one user’s “full-text” search in the title, postco and abstract fields highlighted the need for instructions to show the validation list the user could choose from and an addition to the thesaurus for a cross-reference to the preferred term “full-text.” If there was a validation list/thesaurus the user could use the proper term and successfully retrieve records. The user could easily get frustrated after searching the author field and repeatedly come up zero results. This is another field which a user could benefit from using a validation list. Instead of the user possibly giving up in frustration, the list could give the user immediate feedback and show other author names to explore.

Better control is also necessary when indexing, and assigning terms to articles, so that terms are consistent in syntax and structure. Better instructions regarding the rules of the controlled vocabulary are necessary because many users are not proficient in this area. The instructions should include information on how to search the postco fields using the & / ! for the terms AND, OR, NOT. Including this searching could help facilitate the initial searching process for users unsure of topics or who would like to access a broader scope of articles. The failure of the “indexing” search in the preco and postco fields again showed the need to have a validation list ready for the user to choose the proper term. The instructions could also include a description of the use of the wildcard (\*) feature to retrieve terms such as “index,” “indexer,” and “indexing.” in the title and journal fields. It almost seems that some sort of wildcard filter should be built in to the Preco field, so that all narrowing strings will be retrieved when searching on a term. A filter within the database software would automatically bring up all records with strings beginning with the term entered. This allows the user to find narrower topics within the subject automatically, without any need on their part for additional database instructional knowledge.

The database could be even more useful to users if the order of the fields was changed. The top fields should include the fields used most frequently. A possible new order of the fields could be: abstract, title, journal, author, postco, preco, date, volume. The ID number should be eliminated from the query screen, as it serves no purpose to the searcher and only confuses them. The Pages field could possibly be eliminated from the

query screen based on usage, as it too was found to be confusing. Two field labels should be changed. The title field should be labeled “article title” and journal should be “journal title.” Changing these labels could prevent the possibility of users inputting the title of the journal in the title field.

Concise guidelines can allow users a quick easy glance when using the database. If changes were incorporated into the instructions, users would benefit greatly. The changes to instructions should be kept as brief as possible. Some of the problems users had in retrieving successful results could be solved by including instructions on how to retrieve the list of valid terms. Search results would have increased for the users if instructions included better rules regarding the vocabulary standards and possibly how use Boolean searching, including the wildcard (\*) function.

## REFERENCES

Chowdhury, G. G. (2004). *Introduction to modern information retrieval* (2nd ed.). New York: Neal-Schuman.

David, C. B. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3), 289-299.

Farmer, T. (2005). Making the most of technology: evaluation measures for information retrieval systems. *Arkansas Libraries*, 62(3), 18-22.

Swanson, D. R. (1977). Information retrieval as a trial-and-error process. *Library Quarterly*, 47(2), 128-148.