**INFO 202 SECTION 16: Information Retrieval System Design**

**April 18, 2017**

**Jason Sue**

**Timothy Straus**

**Austin Thompson**

**Project 2: Database Design & Subject Analysis—Bibliographic Database**

**Part A. Design Database**

A1. Descriptor Vocabulary……………………………………………… p2

A2. Statement of Purpose ……………………………………...…...….. p5

A3. Data Structure ……………………………………………………… p6

A4. Rules ……………………………………………………………….. p7

**Part B. Create Content**

B1. Search page URL ……………………………………………...…… p10

B2. Records ………………………………………………………….…. p10

**Part C. Query & Evaluate**

C1. Topics & Queries ………………………………………………….. p15

C2. Evaluation ………………………………………………………….. p17

C3. Reflections …………………………………………………………. p24

**References** …………………………………………………………..…. p29

**Part A: Design Database**

**A1. Descriptor Vocabulary**

academic librarianship

algorithms

archives

Ashby's Law

Blair-Maron experiment

cataloging

children's librarianship

classifications

Claude Shannon

collections

controlled vocabularies

data content

databases

description

design

design process

development

digital content

digital libraries

document representation

document retrieval

documents

effectiveness

evaluation

file structure

films

George Boole

history

home page

indexes

indexing

information

information architecture

information needs

information professionals

information retrieval (IR)

information retrieval systems

information science

information seeking behavior

information technology

interfaces

Internet

law librarianship

lawyers

libraries

Library of Congress Subject Headings (LCSH)

medical librarianship

metadata

Mortimer Taube

natural language

navigation

online public access catalogs (OPAC)

online catalogs

paradigms

precision

public librarianship

query

relevance

Resnikoff-Dolby 30:1 Rule

retrieval

search

search engines

search formulation

standards

storage

subject

systems

technology

television

thesaurus

usability

user experience

user-centered design

users

visual design

websites

young adults

young adult librarianship

**A2. Statement of Purpose**

It can be difficult for MLIS students to find specific information in databases. The research into a specific field can be daunting, as a student may have to search through many articles to find that kind of document they are looking for. For example, if a person is interested in finding articles on information retrieval (IR), they may need to slog through many just to find what they are looking for. For this purpose, MLIS students interested in specific information, in this case IR information, need a database that they can search and always find relevant articles. We aim to provide this sort of database to MLIS students looking to learn more about IR.

Like many other scholarly databases, our database will allow the user to search by both author and title of the article. This will allow them to perform specific searches if they know exactly what they’re looking for. Some people may or may not want to read older articles. Our database allows the user to search by both specific years and by year ranges. The specific year will, obviously, allow the user to find things published within a specific year. Our year range field allows the MLIS student to search by articles from a ten year period ranging from 1970 to present. This way users can choose to only search for material that was published within the last ten years, or look into the history of IR research by searching back to previous decades.

We allow for users to search by the source of the article as well. This will allow them look at our collection of materials based on where the journal comes from. MLIS students may be interested in looking at all of the articles from one journal they deem a reliable source. An abstract field will also be available to search. Using this field, students will be able to use natural language to search the abstracts of the articles.

The descriptor field will be extremely useful to our user base. We have designed a controlled vocabulary that students can use to search the database using this field. The words used for this field were selected after a review of the articles in the database. The vocabulary we use is representative of the contents of each document. Each document has been assigned relevant descriptors, allowing the users to discriminate or aggregate articles more easily.

We also want this database to be able to be readily updated with new information as it becomes available. The field of studying IR is not static, so our vocabulary needs to be flexible enough to incorporate new materials. Our database design and vocabulary will allow new material to be added, though additions to our controlled vocabulary list may be necessary. This database will make research into information retrieval much more straightforward for MLIS students.

**A3. Data Structure**

group29project2  
---------------  
title text comment Title   
author varchar(150) textbox Author   
source varchar(150) textbox Source   
year smallint(6) textbox Year   
abstract text comment Abstract   
descriptor varchar(150) textbox Descriptor   
yearrange varchar(20) list Year Range

**A4. Rules**

1. Field Name: Title

Required?: Yes

Type: Comment (textarea)

Definition: The full title of the articles that are part of the database’s collection.

Entry Rules: The full title must be entered entirely into the system. Use APA format on all titles to ensure consistency. For example, capitalize the first letter of the first word and leave every other word lowercase. If, however, there is a colon somewhere in the title, the word following it must be capitalized also. Example: Historical note: Information retrieval and the future of an illusion.

2. Field Name: Author

Required?: Yes

Type: Textbox

Definition: The full name of the author who may have one or more articles that are part of the database’s collection.

Entry Rules: The full name of the author must be entered entirely into the system. APA format is not required. The last name must come first and the first letter must be capitalized. A comma must be placed afterwards followed by the first name and the first letter must be capitalized. A period must be entered last only if there are no other names provided. If there is a middle or additional name(s), only the first letter must be capitalized immediately followed by a period. Do not spell out middle names. Example: Swanson, Don R.

3. Field Name: Source

Required?: Yes

Type: Textbox

Definition: The full source in which the articles came from that are part of the database’s collection.

Entry Rules: The entire name of the source must be entered completely into the system. Use APA format on all sources to ensure consistency. Capitalize the first word of the title and make sure to capitalize any proper nouns and certain other types of words. Everything else can be lowercase. Add a comma. Type the volume number followed by the issue number in parenthesis. Add a comma. Type the page numbers in numerical order only. Add a period afterwards. Example: Journal of the American Society for Information Science, 50(12), 1043-1050

4. Field Name: Year

Required?: Yes

Type: Textbox

Definition: The full year in which the articles were published that are part of the database’s collection.

Entry Rules: The full year must be entered completely into the system. APA format is not required. Type in the published article’s year as a four digit number. If the published year is unavailable, type in n/a for not available (no capitalization). Example: 1988 or 2008

5. Field Name: Abstract

Required?: Yes

Type: Comment (textarea)

Definition: An empty textarea that makes it convenient for users to input natural language searching for the articles that are part of the database’s collection.

Entry Rules: The entire abstract (from beginning to end) must be entered into the system. The complete abstract must be present, with no alterations in this field.

6. Field Name: Descriptor

Required?: Yes

Type: Textbox

Definition: A textbox for post-coordinate controlled vocabulary that’s appropriate to the subject scope of the articles that are part of the database’s collection.

Entry Rules: Choose from the descriptor list that best matches the topics covered in the article being entered. Separate each descriptor with a comma. There must be a **minimum** of three descriptors and a **maximum** of six descriptors per article.

7. Field Name: Year Range

Required?: Yes

Type: List

Definition: The year range that the article falls into starting from 1970 to present.

Entry Rules: Determine the year of publication of article. If the year falls between 1970 and 1979, select 1970-1979. If the year falls between 1980 and 1989, select 1980-1989. If the year falls between 1990 and 1999, select 1990-1999. If the year falls between 2000 and 2009, select 2000-2009. If the year falls between 2010 and the present year, select 2010-Present.

**Part B. Create Content**

**B1. Search Page URL:** https://libr202.sjsu.edu/webdata\_pro/student/3026/cgi-bin/webdata\_pro.pl?\_cgifunction=user&\_layout=Group29Project2

**B2. Database Records:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **#** | **Title** | **Author** | **Source** | **Year** | **Abstract** | **Descriptor** | **Year Range** |
| 1 | The invisible substrate of information science | Bates, Marcia J. | Journal of the American Society for Information Science, 50(12), 1043-1 | 1999 | The explicit, above-the-water-line paradigm of information... | information, information science, paradigms | 1990-1999 |
| 2 | Vocabulary as a central concept in library and information science | Buckland, Michael | Proceedings of the Third International Conference on Conceptions of Library and Information Science. Dubrovnik, Croatia, 23-26 May 1999 | 1999 | The nature and role of vocabulary in information systems... | controlled vocabularies, digital libraries, indexes, information retrieval (IR), metadata | 1990-1999 |
| 3 | A brief history of information architecture | Resmini, Andrea & Rosati, Luca | Journal ofInformation Architecture, 3(2), 33-46 | 2012 | Information architecture (IA) is a professional practice... | design, history, information, information architecture, information retrieval (IR), information retrieval behavior | 2010-Present |
| 4 | On the shoulders of giants : From Boole to Shannon to Taube: The origins and development of computerized information from the Mid-19th Century to the present | Smith, Elizabeth S. | Information Technology and Libraries, 12(2), 217-226. | 1993 | This article describes the evolvement... | Claude Shannon, George Boole, indexing, information retrieval (IR), Mortimer Taube, technology | 1990-1999 |
| 5 | Historical note: Information retrieval and the future of an Illusion | Swanson, Don R. | Journal of the American Society for Information Science, 39(2), 92 -98 | 1988 | The article presents the historical... | Blair-Maron experiment, history, indexing, information retrieval systems, precision, relevance | 1980-1989 |
| 6 | User-centered design of information systems | Toms, Elaine | M.J. Bates (Ed.), Understanding Information Retrieval Systems: Management, Types, and Standards. Boca Raton, FL : CRC Press. | 2012 | User-Centered Design (UCD) was founded... | design process, development, information, systems, usability, user-centered design | 2010-Present |
| 7 | Design, use, and evaluation of information retrieval systems | Weedman, Judith | Brooke Sheldon & Kenneth Haycock (Eds.), The portable MLIS. Westport, CN: Libraries Unlimited | 2008 | Design is at the core of professional work... | design, evaluation, information professionals, information retrieval (IR) | 2000-2009 |
| 8 | Indexing and access for digital libraries and the internet: Human,  database, and domain factor | Bates, Marcia J. | Journal of the American Society for Information Science, 49(13), 1185-1205. | 1998 | Presents information on a study which looked at indexing... | databases, digital libraries, indexes, Resnikoff-Dolby 30:1 Rule, users | 1990-1999 |
| 9 | The design of browsing and berrypicking techniques for the online search interface | Bates, Marcia J. | Online Review, 13(5), 407-424. | 1989 | First, a new model of searching in online and other information... | design, information retrieval (IR), interfaces, search, systems, user | 1980-1989 |
| 10 | Subject access in online catalogs: A design model | Bates, Marcia J. | Journal of the American Society for Information Science, 37(6), 357- 375. | 1986 | A model based on strikingly different philosophical assumptions... | Ashby's Law, design, Library of Congress Subject Headings (LCSH), online public access catalogs (OPAC) | 1980-1989 |
| 11 | Library data in a modern context | Coyle, Karen | Library Technology Reports, 46(1), 5-13. | 2010 | The article examines the state of library data... | cataloging, Internet, metadata, online public access catalogs (OPAC), search, users | 2010-Present |
| 12 | Metadata for all: Descriptive standards and metadata sharing across libraries, archives, and museums | Elings, Mary W. & Waibel, Gunter | First Monday, 12(3). | 2007 | Integrating digital content from libraries, archives and museums represents... | data content, development, digital content, indexing, metadata, standards | 2000-2009 |
| 13 | Issues in the development of a thesaurus for patients' chief complaints in the emergency department | Haas, Stephanie W. & Travers, Debbie A | 67th Proceedings of the ASIS&T annual meeting, 41, 411-417. | 2004 | When a patient visits the Emergency Department (ED)... | controlled vocabularies, design process, development, medical librarianship, natural language, thesaurus | 2000-2009 |
| 14 | Representing and accessing information in the stockshot database at the National Film Board of Canada | Turner, James | Representing and accessing information in the stockshot database at the National Film Board of Canada. The Canadian Journal of Information Science - | 1990 | This article discusses the challenges of representing visual material... | information retrieval (IR), interfaces, search formulation, storage, user experience, visual design | 1990-1999 |
| 15 | Design thinking: A useful myth and rethinking design thinking | Norman, Donald A. | <http://www.jnd.org/dn.mss/design_thinking_a_u.html.> | 2010 | What is design thinking? It means stepping back from the immediate issue... | design, design process, effectiveness | 2010-Present |
| 16 | Learning from librarians and teens about YA library spaces | Agosto, Denise E., Kuhlmann, L. Meghann, Pacheco Bell, J., & Bernier, Anthony | Public Libraries, 53(3), 24-28 | 2014 | The article discusses the results of the empirical study... | evaluation, public librarianship, user experience, user-centered design, young adults, young adult librarianship | 2010-Present |
| 17 | Information search tactics | Bates, Marcia J. | Journal of the American Society for Information Science, 30(4), 205-214. | 1979 | As part of the study of human information search strategy... | databases, information seeking behavior, search formulation | 1970-1979 |
| 18 | Relevance for browsing, relevance for searching | Bodoff, David | Journal of the American Society for Information Science and Technology 57(1), 69-86. | 2006 | The concept of relevance has received a great deal of theoretical attention... | algorithms, retrieval, relevance, search, users | 2000-2009 |
| 19 | Manipulating search engine algorithms: the case of Google | Bar-Ilan, Judit | Journal of Information, Communication and Ethics in Society 5(2/3), 155-166. | 2007 | Purpose - To investigate how search engine users manipulate the rankings of search... | algorithms, databases, Internet, query, search engines, search formulation | 2000-2009 |
| 20 | An evaluation of retrieval effectiveness for a full-text document-retrieval system | Maron, M. E. & Blair, David C. | Communications of the ACM, 28(3), 289-299. | 1985 | An evaluation of a large, operational full-text document-retrieval system... | databases, document retrieval, lawyers | 1980-1989 |
| 21 | Homepage real estate allocation. | Nielsen, Jakob | <http://www.nngroup.com/articles/homepage-real-estate-allocation/> | 2013 | Websites spend too little homepage screen space... | design, information, Internet, navigation, user experience | 2010-Present |

**Part C. Query & Evaluate**

**C1. Topics & Queries**

**Topic 1: thesaurus development and standards.**

**Queries in descriptor field:**

1. thesaurus OR controlled vocabularies AND standards. Retrieved 1 record. The record number was: 13

2. thesaurus OR controlled vocabularies. Retrieved 2 records. The record numbers were: 2, 13

**Queries in abstract field:**

1. thesaurus OR controlled vocabularies AND standards. Retrieved 2 records. The record IDs were: 10, 13

2. thesaurus OR controlled vocabularies. Retrieved 2 records. The record numbers were: 10, 13

**Topic 2: The application of controlled vocabularies in medical librarianship.**

**Queries in descriptor field:**

1. medical librarianship AND controlled vocabularies. Retrieved 1 record. The record number was: 13

2. controlled vocabularies OR thesaurus AND medical librarianship. 2, 13

**Queries in abstract field:**

1. medical librarianship AND controlled vocabularies. Retrieved no records.

2. controlled vocabularies OR thesaurus AND medical librarianship. Retrieved no records.

**Topic 3: I’d like to know about databases used in libraries**

**Queries in descriptor field:**

1. databases AND libraries. Retrieved 1 record. The record numbers was 8

2. catalogs OR databases. Retrieved 6 records. The record numbers were 8, 10, 11, 16, 19, 20

**Queries in abstract field:**

1. databases AND libraries. Retrieved no records.

2. catalogs OR databases. Retrieved 1 record. The record number was 13

**Topic 4: I’d like to learn about the design of information retrieval systems**

**Queries in descriptor field:**

1. information retrieval AND design. Retrieved 4 records. The record numbers were 3, 7, 9, 13

2. information retrieval OR systems AND design. Retrieved 8 records. The record numbers were 2, 3, 4, 5, 6, 7, 13

**Queries in abstract field:**

1. Information retrieval AND design. Retrieved 3 records. The record numbers were 7, 9, 19

2. Information retrieval OR systems AND design. Retrieved 8 records. The record numbers were 2, 5, 6, 7, 14, 16, 19

**Topic 5: The study of user searches in IR systems**

**Queries in descriptor field:**

1. search AND users. Retrieved 1 record. The record number was 9

2.information retrieval systems AND search AND users. Retrieved no records.

**Queries in abstract field:**

1. search AND users. Retrieved 2 records. The record numbers were 11, 18

2.information retrieval systems AND search AND users. Retrieved 6 records. The record numbers were 8, 9, 11, 14, 17, 18

**Topic 6: Information Retrieval Indexes**

**Queries in descriptor field:**

1.indexing AND information retrieval (IR). Retrieved one record. The record number was 4

2.information retrieval (IR) OR information retrieval systems AND indexing OR indexes. Retrieved 8 records. The record numbers were 2, 3, 4, 5, 7, 8, 9, 14

**Queries in abstract field:**

1.indexing AND information retrieval (IR). Retrieved 1 record. The record number was 5

2.information retrieval (IR) OR information retrieval systems AND indexing OR indexes. Retrieved 2 records. The record numbers were 2 and 5

**C2. Evaluation**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Topic** | **Relevant articles to the Topic (article #’s)** | **Queries** | **Relevant articles retrieved with Abstract field search (article #’s)** | **Relevant articles retrieved with Descriptor field search (article #’s)** | **Recall** | **Precision** | **Eff** |
| 1 | 10  12  13 | thesaurus OR controlled vocabularies AND standards | 10, 13 | 13 | Abstract:  0.6667  Descriptor  0.3333 | Abstract:  1.0000  Descriptor1.0000 | Abstract:  0.4730  Descriptor:  0.3333 |
| thesaurus OR controlled vocabularies | 10, 13 | 2, 13 | Abstract:  0.6667  Descriptor:0.3333 | Abstract:  1.0000  Descriptor 0.5000 | Abstract:  0.4730  Descriptor:  0.6667 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Topic** | **Relevant articles to the Topic (article #’s)** | **Queries** | **Relevant articles retrieved with Abstract field search (article #’s)** | **Relevant articles retrieved with Descriptor field search (article #’s)** | **Recall** | **Precision** | **Eff** |
| 2 | 13 | medical librarianship AND controlled vocabularies |  | 13 | Abstract:  0.0000  Descriptor1.0000 | Abstract:  0.0000  Descriptor1.0000 | Abstract:  0.0000  Descriptor:  1.0000 |
| controlled vocabularies OR thesaurus AND medical librarianship |  | 2, 13 | Abstract:  0.0000  Descriptor1.0000 | Abstract:  0.0000  Descripto:0.5000 | Abstract:  0.0000  Descriptor:  0.6464 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Topic** | **Relevant articles to the Topic (article #’s)** | **Queries** | **Relevant articles retrieved with Abstract field search (article #’s)** | **Relevant articles retrieved with Descriptor field search (article #’s)** | **Recall** | **Precision** | **Eff** |
| 3 | 8  10  11  12 | databases AND libraries | 0 | 8 | Abstract:  0.0000  Descriptor:  0.25 | Abstract:  0.0000  Descripto: 1.000 | Abstract:  0.0000  Descriptor:  0.4697 |
| catalogs  AND  databases | 8, 10, 11 |  | Abstract:  0.0000  Descriptor:  0.7500 | Abstract:  0.0000  Descripto:0.5000 | Abstract:  0.000  Descriptor:  0.6047 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Topic** | **Relevant articles to the Topic (article #’s)** | **Queries** | **Relevant articles retrieved with Abstract field search (article #’s)** | **Relevant articles retrieved with Descriptor field search (article #’s)** | **Recall** | **Precision** | **Eff** |
| 4 | 3  5  7  9  13  14 | Information retrieval AND design | 7, 9 | 3, 7, 9, 13 | Abstract:  0.3333  Descriptor:  0.6667 | Abstract:  0.6667  Descriptor1.0000 | Abstract:  0.4730  Descriptor:  0.7643 |
| Information retrieval OR systems AND design | 5, 7, 9 | 3, 5, 7, 9, 13 | Abstract:  0.5000  Descriptor:  0.8333 | Abstract:  0.3750  Descripto:0.6250 | Abstract:  0.4340  Descriptor:  0.7098 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Topic** | **Relevant articles to the Topic (article #’s)** | **Queries** | **Relevant articles retrieved with Abstract field search (article #’s)** | **Relevant articles retrieved with Descriptor field search (article #’s)** | **Recall** | **Precision** | **Eff** |
| 5 | 6  7  9 | search AND users | 9, 19 | 11, 18 | Abstract:  0.2500  Descriptor:  0.0000 | Abstract:  1.0000  Descriptor  0.0000 | Abstract:  0.4297  Descriptor:  0.0000 |
| search AND information OR users | 1, 2, 6, 9, 16, 17, 19, 21 | 8, 9, 11, 14, 17, 18 | Abstract:  0.6666  Descriptor:  0.3333 | Abstract:  1.0000  Descriptor  1.0000 | Abstract:  0.7643  Descriptor:  0.5286 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Topic** | **Relevant articles to the Topic (article #’s)** | **Queries** | **Relevant articles retrieved with Abstract field search (article #’s)** | **Relevant articles retrieved with Descriptor field search (article #’s)** | **Recall** | **Precision** | **Eff** |
| 6 | 2  4  5  8 | indexing AND information retrieval (IR) | 5 | 4 | Abstract:  0.2500  Descriptor:  0.2500 | Abstract:  1.0000  Descriptor  1.0000 | Abstract:  0.4297  Descriptor: |
| information retrieval (IR) OR information retrieval systems AND indexing OR indexes | 2, 5 | 2, 3, 4, 5, 7, 8, 9, 14 | Abstract:  .5000  Descriptor:  1.0000 | Abstract:  1.0000  Descriptor1.0000 | Abstract:  0.6464  Descriptor:  1.000 |

**Takeaways:**

Jason Sue selected topic 1: thesaurus development and standards and topic 2: the application of controlled vocabularies in medical librarianship. For the first topic, records 10. 12. and 13 were deemed to be relevant. Because thesaurus and controlled vocabulary were determined to be near synonyms for the purposes of topic 1, Topic 1’s first query was thesaurus OR controlled vocabularies AND standards. When searching the descriptor field it only retrieved document surrogate 13. This resulted in a recall of 0.3333, a precision of 1.000, and an effectiveness of 0.4730. When searching the natural language field abstract, recall increased and records 10 and 13 were retrieved. This resulted in a recall of 0.667, a precision of 1.0000 and and an effectiveness of 0.4730 . Topic 1’s second query was thesaurus OR controlled vocabularies. By leaving out a descriptor in the query Jason Sue hoped to increase recall. When searching the descriptor field it only retrieved document surrogate 2, and 13. Unfortunately, record 2 was not relevant. This resulted in a recall of 0.3333, a precision of 0.5000, and an effectiveness of 0.6667. When searching the natural language field abstract, recall increased and records 10 and 13. Were retrieved. This resulted in a recall of 0.667, a precision of 1.0000 and and an effectiveness of 0.4730 . For topic 2 the first query was medical librarianship AND controlled vocabularies. This resulted in a recall of 1.000, a precision of 1.0000 and and an effectiveness of 1.000 . The second query was medical librarianship AND controlled vocabularies. This resulted in a recall of 1.000, a precision of .0,5000 and and an effectiveness of 0.6464 . No results were retrieved in the abstract.

Timothy’s topics were 3 and 4 with the third topic proving significantly less successful than the fourth. No relevant results turned up at all in the natural language search in topic 3, but what is really interesting took place in the descriptor searches. During the first search for databases AND libraries, only one relevant document was called, but in the second using catalogs AND databases, three relevant files were recalled including the one found in the first search. Since it was predetermined that these documents were relevant to the search topic, it seems that leaving out libraries as a descriptor may have been a mistake. Careful consideration is needed for each article to ensure that it has all of the most important descriptors that may be used to locate it.

Austin’s topics were numbers 5 and 6. The fifth topic proved to be less successful than the sixth topic. There was even an instance where no results were found during an abstract search. For the sixth topic, the results were very relevant. Every single one of the predetermined relevant articles were found in both the abstract and descriptor searches. This may have had something to do with the topics. The fifth topic had a lot of different subjects that needed to be together to fit the information search. The keywords were information retrieval (IR), search, systems, and users. However, the sixth topic only had two subjects to focus on. The two keywords were both indexes and information retrieval (IR). In order to improve retrieved results that are relevant in the fifth topic, the descriptor terms must be chosen wisely. For example, the sixth and seventh articles in the database should have the descriptor vocabulary term “search” to make it easier to find for someone who wants to research “study of user search behavior in IR systems.” Even though the articles were more heavily focused on design, they did discuss the use or the “search” aspect in the design process of IR systems. Austin learned from this project a very important key takeaway: poor labeling of documents will produce little or no results to the user.

**C3. Reflections**

**Group Reflections**

The work was divided between the three of us as equally as we could make it. As with project 1, Austin took on making the database and filling it and Tim wrote the SOP and pulled together the reflections in Part 3C. Jason aggregated the vocabulary terms and posted them to the discussion thread. Each member was responsible for generating vocabulary terms for six of the articles with Austin taking the seventh and everyone sharing responsibility for the Bates article.

The most difficult part of this project was selecting the subject terms for the vocabulary. Relevance is always a dictating factor in determining terms, but determining the relevance of the terms can be difficult. It is easy to select terms that look important, and use that. In reality, that term may not have that much overall significance to the article. What did work well was looking toward the more general themes of an article and picking terms that more generally applied to it. Using that method, terms more easily cover the larger body of articles and are more relevant.

Determining the main subject of an article can be difficult. When one is looking for words to use as terms for an article, it is easy to get lost looking for buzzwords more than looking at the overall meaning of the article. Skimming makes it hard to read for content, but can also be extremely useful for finding words that are frequently repeated. Another way to help make sure one is on the right track is to review the abstract of the article. Abstracts are designed to represent the article in a nutshell. Using the abstract as a guide, one can check their terms in order to make sure the main subject of the article is conveyed in the vocabulary.

There are some improvements that we think we could make to our database to improve the user experience. For example, we discussed adding a field that would contain the APA citation for the articles that we have in our database. Ultimately we decided it was too time consuming for this project, but it would be a nice feature. Another change that could be made to the database would be to hide the year field entirely from the user. It seems unlikely that people would search by a specific year and would more likely use the year range field, so the year field adds clutter to the front end of the database.

We found that our descriptors frequently performed more reliably than the same search in the abstract field. In a number of cases throughout our queries, we would get at least one relevant article from the descriptors, but irrelevant results or nothing at all from the abstract field. There are some topics, namely topic 1, where abstract searching performed better than the descriptor searches. That tells us that our descriptor terms may not be effective as at generating results for that particular topic. Either more careful assignment of descriptors or finding better descriptors for related documents would help to fix the issue.

**Timothy Straus**

The main takeaway from this assignment for me was in regards to aboutness. Throughout this semester, I have been brought back to a quote from earlier in our course, “Simply put, relevant results are those which are interesting and useful to the user” (Morville, 2005, p. 49). Aboutness is all about relevance. When we select a term for a controlled vocabulary, we are selecting it because it is about the subject of an article that it will represent. When we select a term, we want to make sure that the user will use it when searching on the subject, or else the term will not produce relevant results. Without thinking about aboutness, we cannot make a database where there are findable documents. Findability depends on our user’s ability to aggregate or discriminate documents within our database. If the terms are not true to the item they are assigned to, proper aggregation and discrimination are impossible and the item is no longer findable. Not only does the aboutness of the terms matter, but also the amount of terms assigned to a document. The more terms that are used the harder it is to discriminate against an article. The less terms that are used the more difficult it becomes to aggregate with other similar items. From assignment of terms to the aboutness of them, findability and relevant results suffer if any aspect of the controlled vocabulary design process is neglected.

**Austin Thompson**

With every system designed strictly for information retrieval, there must be the same end goal for every single one: relevance. If the IR system doesn’t give its users what they want, the system completely becomes irrelevant to them. Why bother creating an IR system if it doesn’t work? It can be tough to make sure every type of information need is met since every user is different from other users in terms of what they’re looking for. A great example is the depth and focus of a single document. After all, “a document may be deep and focused, or it may be a general overview, but we cannot say which depth level is better for relevance in general, without reference to the user need; it depends on what the person needs” (Bodoff, 2006, p. 71). In other words, the value is subjective to every user who reads the same document. The design of IR systems shouldn’t be taken lightly. Design should have an equal separation of thinking and doing. Forcing the user to do a lot (physical) or think a lot (cognitive) will cause the user to get frustrated or stressed out (affective) and abandon the system entirely. From the perspective of a system designer, “when all of the elements, procedures, and objects work in harmony, the system almost ‘disappears’ as the user is ‘one’ with the visible components” (Toms, 2012, p. 69). First, the ease of use should be a big focus in user-centered design. Second, the IR system itself should only be judged by two factors: “how well it captures relevant documents, and how will it rejects the irrelevant” (Swanson, 1988, p. 92). This course (and specifically this project) has taught me findability only works well if the design works well in information retrieval system design.

**Jason Sue**

Personally, communication has been really difficult for me because throughout the whole assignment my notes referred to the articles 1-49 and Austin and Tim used the condensed 1-21. Whenever they used numbers, I had to to translate what they were speaking about and when posting my results I again had to translate my results so that they corresponded to the group’s numbering system. It would have been prudent of me to accept the condensed numbering system once the articles had been selected to avoid creating unnecessary barrier to communication as well as avoiding a great deal of unnecessary work.

During the course of the discussion, I realized that my thought process is fundamentally differently than Austin Thompson or Timothy Straus. They are both place able to see the big picture much more readily than I, which makes both of them great leaders. I on the other hand take notice of the details and am more likely to bring up small discrepancies in the results or formating. It normally takes a couple minutes for me to identify broader themes and ideas from the minutia.

Much of the group’s difficulty had origins in vocabulary selection which was primarily my responsibility. Page 4 from Lecture 7: Evaluation identifies two sources of concern when it comes to a controlled vocabulary. The first is if there are no terms in the controlled vocabulary to describe the subject. For the most part, the controlled vocabulary is able to describe the aboutness of each article. The second is if the indexer left out a relevant term or if they assigned a term that only describes a minor part of the document. Unfortunately, the vocabulary developed under my direction made tern assignment to be unnecessarily problematic. In short, I had too many closely-related terms in the controlled vocabulary., I wanted to be able to be able to discriminate between and information retrieval systems and the more general information retrieval. From the instructions from Vocabulary Design Exercise on slide 22, I remembered how Judy Weedman, Ph.D. accepted both academic libraries and academic librarians as two different terms and suggested that a related term relationship be used to established between them. Under the impression that I was selecting a database for a large number of articles, where students of information science would be able to use the controlled vocabulary to create precisely tailored queries to meet their information needs. Unfortunately, this backfired. In particular, I overlooked the amount of trouble that indexers would face when they were limited to only six descriptor terms.

**References**

Bodoff, D. (2006). Relevance for browsing, relevance for searching. *Journal of the American Society for Information Science and Technology, 5*7(1), 69-86.

Morville, P. (2005). *Ambient Findability*. Seabstopol, CA: O’Reilly.

Swanson, D. R. (1988). Historical note: Information retrieval and the future of an Illusion. *Journal of the American Society for Information Science, 39*(2), 92 -98.

Toms, E. (2012). User-centered design of information systems. In M.J. Bates (Ed.), *Understanding Information Retrieval Systems: Management, Types, and Standards.* Boca Raton, FL:CRC Press.

Tucker, V. M. (2014) Lecture 7: Evaluation [PDF]

Weedman, J. (2014). Vocabulary design exercise: Getting started [PDF].