IN Harmony: Sheet Music from Indiana
Query Logs Analysis Evaluation Plan

Indiana University (IU), the Indiana State Library (ISL), the Indiana State Museum (ISM), and the Indiana Historical Society (IHS) have received a grant from the Institute of Museum and Library Services (IMLS) to provide electronic access to Indiana-related sheet music from each of the institutions' collections. Approximately 10,000 pieces of sheet music will be available online as a result of this grant activity. One of the primary goals for this project is to provide robust, consistent browse and search access across all participants' collections or within a particular collection. As a result, cataloging guidelines and tools for sheet music description will be developed to aide the project partners in a) cataloging sheet music not yet described in their respective collections; and b) map existing cataloging records to a format that will facilitate cross-collection searching.

Background

Prior to receiving the IMLS grant, the Indiana University Digital Library Program created a sheet music website, Indiana University Sheet Music [http://www.dlib.indiana.edu/collections/sheetmusic/] featuring digitized sheet music from the Lilly Library and has contributed to a collaborative sheet music project known as the Sheet Music Consortium [http://digital.library.ucla.edu/sheetmusic/][1]. We have learned as a result that users of sheet music have unique discovery needs especially in regard to subject access. Existing research in this area has confirmed that subject access is not approached by users in a uniform way; subject searches typically include topic, form, genre, style and geographic terms thereby potentially complicating access (Cunningham, Reeves & Britland, 2003; Pachet & Cazaly, 2000; Fabbri, 1999). While the research literature explores aspects of subject categorization, much of the existing research seems to lack usage-based analysis.

Purpose of Study

In order to better understand how sheet music is browsed and searched, we plan to evaluate a subset of the query logs captured by the Indiana University Sheet Music Collection (housed locally) and the Sheet Music Consortium (housed at University of California, Los Angeles) websites.

We are particularly interested in learning:

- How often users conduct a browse, search or advanced search for sheet music
- How often users conduct known-item (specific) versus unknown-item (general) searching
- What kinds of searches are being conducted (keyword, title, name, subject, etc.)

[1] The Sheet Music Consortium project is based in the University of California, Los Angeles, where the website is also housed. Contributing partners also included Johns Hopkins and Indiana University. The website was launched summer 2003.

- What kinds of subject-related queries are being entered by users (e.g. topical, genre, style, etc.)

Answering these questions will help us not only design browse and search features that meet user needs for the IN Harmony website, but will also provide guidance in how sheet music should be cataloged. Because of the lack of user-based research in this area, our findings may benefit others designing online sheet music and possibly general music web resources for students, scholars or the community at-large.

Methodology: Data Gathering, Processing and Analysis

The existing log data will come from two sheet-music related sources: the Indiana University Sheet Music and the Sheet Music Consortium websites. Because the Indiana University Digital Library Program (IUDLP) is affiliated with both of these projects, log data can be easily obtained. Stephen Davison, Head of the UCLA Digital Library Program, has agreed to share with the IUDLP the Sheet Music Consortium logs.

Only a subset of the logs will be analyzed. A tentative 6-month period of log activity has been identified for analysis: June – November 2004. The log data consists of:

- IP Addresses (which we will not be utilizing; IU-only)
- Timestamps
- Types of searches (browse, advanced, simple)
- Queries
- Fields selected (keyword, name, title, etc.)
- Boolean or other advanced operators

Personal identifiers are not included in this log data.

Based on the Indiana University Sheet Music logs, an average of 37,904 queries are posted to the website per month.[2] The combination of UCLA and IU log data will increase that number considerably. Random selection of records for analysis will occur to meet project deadlines. Our goal is to analyze approximately 5,000 records.

Data Processing

In order to makes sense of the log data, processing for analysis will have to occur. For data residing in our servers, processing documentation (see Appendix A for instructions) has been completed and distributed to our programmer. The date processing output consists of two types of formats: plain text file with special formatting and a CSV file used for importing data into spreadsheets (see Appendix B for examples of each).

Once sample data is received from UCLA, processing documentation will be completed for data reformatting for the Sheet Music Consortium logs in a manner similar to that used for the IU logs.

[2] Based on June 23-November 23, 2004 log data.

Data Analysis

Microsoft's Excel spreadsheet program will be primarily used for analysis. Data can be easily sorted, coded and displayed in various views including graphical representations.

General goals for data analysis have been identified (see list below). Other goals and criteria may surface upon closer inspection of the processed data:

- Determine relative frequency of browse, search and advanced searches conducted
- Compare number of known-item to unknown-item queries
- Sort the queries into subject-related search strings as well as by other fields such as creator, title and other identifiable access points for further evaluation
    - Determine further categories for subject-related search strings (topical, form, genre, style, temporal, geographic, etc.)
- Compare subject data to a set of sample cataloged records from the partnering institutions to determine overlap of user search terms with controlled vocabularies used in subject description (e.g. Library of Congress Subject Headings, Art and Architecture Thesaurus, etc.)

References

Cunningham, S.J., Reeves, N. & Britland, Matthew. (2003). An ethnographic study of music information seeking: Implications for the design of a music digital library. *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries, Houston,* 5-16.

Fabbri, F. (1999). Browsing music spaces: categories and the musical mind, presented at the 3rd Triennial British Musicological Societies Conference (Guildford, UK, 1991), available at http://www.theblackbook.net/acad/others/fabbri990717.pdf.

Pachet, F. & Cazaly, D. (2000). A taxonomy of musical genres. *Proceedings of Content-Based Multimedia Information Access Conference (RIAO), Paris*.

Appendix A: Processing Documentation for Indiana University Sheet Music Collection Logs

Queries posed to the IU Sheet Music Collection web site (http://www.letrs.indiana.edu/web/s/sheetmusic/) need to be captured for analysis. The required information is entirely contained in the URL querystring (see examples below).

Every URL request is captured in apache logs on Algernon (http://algernon.dlib.indiana.edu/): **/opt/apache/logs/**. The current log file is *access_log* and the archive files are *access_log.1.gz* thru *access_log5.gz*.

These logs capture access information for multiple collections not just the sheet music collection. The name-value pairs highlighted below are discussed in more detail following the example URLs.

Example URLs for IU Sheet Music

The IU Sheet Music Collection currently contains records from two collections: Starr and Devincent. These can be searched individually or together. Sample URLs for a browse, simple and advanced search are included below. Each essential name-value pair is highlighted within the querystring.

Browse

The browse actually functions like a simple search. The name browse searches within the "names" field, and the title browse searches within the "title" field. The query string for the browse is constructed differently than the simple search. We can differentiate between a browse and simple search based on the querystring syntax.

Name browse:

http://www.letrs.indiana.edu/cgi/b/bib/bib-idx?&**g=sheetmusic**&**c=starr&c=devincent**&view=reslist&**type=simple**&**q1=Aaron%2C%20Jack**&**rgn1=names**

Title browse:

http://www.letrs.indiana.edu/cgi/b/bib/bib-idx?&**g=sheetmusic**&**c=starr&c=devincent**&view=reslist&**type=simple**&**q1=Galop%20/%20/%20composed%20by%20L.%20Streabbog**&**rgn1=title**

Simple Search

1 term query:

http://www.letrs.indiana.edu/cgi/b/bib/bib-idx?**type=simple**&**c=devincent&c=starr**&sid=365b360706dfd904ae463935902f4854&xc=1&**g=sheetmusic**&**q1=waltz**&**rgn1=entire+record**

2 or more term query (treated as phrase):

http://www.letrs.indiana.edu/cgi/b/bib/bib-idx?**type=simple**&**c=devincent&c=starr**&sid=365b360706dfd904ae463935902f4854&xc=1&**g=sheetmusic**&**q1=blue+moon**&**rgn1=entire+record**

Advanced Search

Multiple terms/Boolean query with date:

http://www.letrs.indiana.edu/cgi/b/bib/bib-idx?**type=boolean**&**c=devincent&c=starr**&sid=11686a633b1cb07228d314dbed305799&xc=1&**g=sheetmusic**&**q1=love**&**rgn1=subject**&**op2**=**Or**&**q2=marriage**&**rgn2=subject**&**op3**=**And**&**q3=new+york**&**rgn3=entire+record**&**date1=1800**&**date2=2004**

Name-Value Pairs

g=sheetmusic	The collection website name follows the "g". This is important for selecting sheetmusic-specific URLs from the logs, but it is not the only indicator for selecting querystrings to be parsed.
c=starr&c=devincent	The sheet music collection name follows the "c". They can together (c=starr&c=devincent) or individually (c=starr).
type=simple **type=boolean**	The type contains the kinds of search. "Simple" can be a simple search or browse and "Boolean" is an advanced search.
q1=love&**rgn1=subject**	The query terms entered follows "q1", "q2" and/or "q3". More than one term can be included in one query (e.g. **q1=blue+moon**). The field in which that terms was searched appears immediately after the "q#" value: **rgn1=subject**.
op2=Or **op3=And**	For advanced searches, the Boolean operator selected follows either "op2" or "op3".
date1=1800&**date2=2004**	For advanced searches, date ranges can be used. A range is always required so "date1" and "date2" will always appear in the querystring.

Log Examples

Example log entries are included below. They include an IP number, timestamp, form method, URL and HTTP header information.

Browse

129.79.37.99 - - [09/Nov/2004:10:57:59 -0500] "GET /cgi/b/bib/bib-idx?&g=sheetmusic&c=starr&c=devincent&view=reslist&type=simple&q1=Aaron%2C%20Jack&rgn1=names HTTP/1.1" 200 10320

129.79.37.99 - - [09/Nov/2004:11:07:01 -0500] "GET /cgi/b/bib/bib-idx?&g=sheetmusic&c=starr&c=devincent&view=reslist&type=simple&q1=Galop%20/%20/%20composed%20by%20L.%20Streabbog&rgn1=title HTTP/1.1" 200 10484

Simple Search

129.79.37.99 - - [09/Nov/2004:10:59:31 -0500] "GET /cgi/b/bib/bib-idx?type=simple&c=devincent&c=starr&sid=365b360706dfd904ae463935902f4854&xc=1&g=sheetmusic&q1=waltz&rgn1=entire+record HTTP/1.1" 200 45799

Advanced Search

211.246.143.41 - - [09/Nov/2004:16:20:35 -0500] "GET /cgi/b/bib/bib-idx?type=boolean&c=devincent&c=starr&sid=b2f72e9ac7f3f54944d33aea71d72f8c&xc=1&g=sheetmusic&q1=Radetzky+Marsch+&rgn1=subject&op2=And&q2=Johann+Strauss&rgn2=names&op3=And&q3=&rgn3=entire+record&date1=1800&date2=2004 HTTP/1.1" 200 6753

Boolean with empty q2 and q3 strings:

212.5.70.188 - - [09/Nov/2004:13:13:56 -0500] "GET /cgi/b/bib/bib-idx?type=boolean&c=devincent&c=starr&sid=663f2e0c4bb549e8d5bfd20e2cc654a8&xc=1&g=sheetmusic&q1=Mozart&rgn1=entire+record&op2=And&q2=&rgn2=entire+record&op3=And&q3=&rgn3=entire+record&date1=1800&date2=2004 HTTP/1.0" 200 14143

URLs Not To Be Parsed

Several URLs also contain the **g=sheetmusic** name-value pair that do not need to be parsed. Some examples are included below. Others may need to be identified.

Sort/Submit (from the results page):

80.130.129.203 - - [09/Nov/2004:12:53:59 -0500] "GET /cgi/b/bib/bib-idx?type=simple&c=devincent&c=starr&g=sheetmusic&sid=429c80ec0b5115fa59cab4f9937fbb8f&xc=1&**Submit=search&sort=A-Z**&q1=the+moldau&rgn1=entire+record HTTP/1.1" 200 6657

Results of Page views:

80.130.129.203 - - [09/Nov/2004:12:55:31 -0500] "GET /cgi/b/bib/bib idx?c=devincent;g=sheetmusic;cc=devincent;xc=1;sid=429c80ec0b5115fa59cab4f9937fbb8f;rgn1=id;q=LL-SDV-217016;type=simple;**view=reslistlong**;**fmt=long** HTTP/1.1" 200 3015

129.79.37.99 - - [09/Nov/2004:11:03:10 -0500] "GET /cgi/b/bib/bib-idx?c=devincent;c=starr;g=sheetmusic;cc=ALLSELECTED;xc=1;sid=365b360706dfd904ae463935902f4854;q1=love;q2=marriage;q3=new%20york;op2=Or;op3=And;date1=1800;date2=2004;**page=boolean** HTTP/1.1" 200 9015

Parse Output

The logs need to be parsed in sequential steps. The goal is to extract the necessary data from the logs as identified above. The script that does this should be generalizable enough to use for other DLXS-based collections for which we may need query analysis. For this particular analysis, the data needs to be ported to an Excel spreadsheet (in CSV or other delimited format).

Example: 211.246.143.41 - - [09/Nov/2004:16:20:35 -0500] "GET /cgi/b/bib/bib-idx?**type=boolean&c=devincent&c=starr**&sid=b2f72e9ac7f3f54944d33aea71d72f8c&xc=1&**g=sheetmusic&q1=Radetzky+Marsch**+&**rgn1=subject&op2=And&q2=Johann+Strauss&rgn2=names&op3=And&q3=&rgn3=entire+record&date1=1800&date2=2004** HTTP/1.1" 200 6753

Output 1: Labeled (Text File)

Website:	sheetmusic (from g=)
IP Number:	211.246.143.41
Timestamp:	09/Nov/2004:16:20:35 -0500
Year:	2004
Month:	11
Day:	09
Collection:	devincente (from c=)
Collection:	starr (from c=)
Type:	boolean [(browse) | simple | boolean]
Query:	subject: Radetzky Marsch And names: Johann Strauss And 1800-2004 [term = field (And | Or) date1-date2] (from q1-q3= + rgn1-rgn3 + op2-op3 + date 1-date 2)

Output2: CSV (Excel File)

Website
IP Number
Timestamp
Year
Month
Day
Collection
Collection
Type
Query 1
Field 1
Boolean 1
Query 2
Field 2
Boolean 2
Query 3
Field 3
Date 1
Date 2

Rules:

- Separate timestamp into year, month, and day. Map alpha month to numeric form (*e.g. Nov = 11*).
- Create two labels, one for each collection searched. If only one searched, then only display one label.
- Indicate search type. If possible distinguish between "browse" and "simple".
- Field searched (rgn1-3) should appear before term, separated by a colon (*e.g. entirerecord: blue sky ; subject: love*)
- For the Output 1, if q2-3 = null, then don't print q2 + rgn2, q3 + rgn3. For Output 2, represent all NULL values (e.*g. indicate blank fields with Nulls, Query2 = NULL*).
- Remove/Map URL encoding and HEX values between terms (*e.g. q1=Aaron%2C%20Jack&rgn1=names should be processed as Aaron, Jack*). See: http://www.blooberry.com/indexdot/html/topics/urlencoding.htm for more info: %20 = Space and 2C = comma.

Appendix B: Examples of Processed Data

CSV Format for Excel

23/Jun/2004:12:46:09 -0500	browse	Ha, ha, ha, ha, ho, ho, ho, ha, ha, ho, ho, ho	title			
23/Jun/2004:12:46:37 -0500	browse	Hallelujah, Hallelujah ev'rybody cried, Johnny swelled	title			
23/Jun/2004:14:21:18 -0500	simple	Champagne's+delicious+bubbles+	entire+record			
23/Jun/2004:16:40:28 -0500	boolean	ravel	entire+record	And	hebraiques	entire+record

Text File, Labeled Format

Website: sheetmusic
IP Number: 198.4.83.52
Timestamp: 23/Jun/2004:12:46:09 -0500
Year: 2004
Month: 6
Day: 23
Collection: starr
Collection: devincent
Type: browse
Query: title: Ha, ha, ha, ha, ho, ho, ho, ha, ha, ho, ho, ho

Website: sheetmusic
IP Number: 198.4.83.52
Timestamp: 23/Jun/2004:12:46:37 -0500
Year: 2004
Month: 6
Day: 23
Collection: starr
Collection: devincent
Type: browse
Query: title: Hallelujah, Hallelujah ev'rybody cried, Johnny swelled