Chopin Early Editions: Construction and Usage of Online Digital Scores

Tod A. Olson
The University of Chicago Library
1100 E. 57th Street, JRL 220
Chicago, IL 60637
tod@uchicago.edu

J. Stephen Downie
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 East Daniel St.
Champaign, IL 61820
jdownie@uiuc.edu

Abstract

The University of Chicago Library has digitized a collection of 19th century music scores. The online collection is generated programmatically from the scanned images and human-created descriptive and structural metadata, encoded as METS objects, and delivered using the Greenstone Digital Library software. Use statistics are analyzed and possible future directions for the collection are discussed.

1 Introduction

The University of Chicago Library is home to an active collection of over 400 first and early edition music scores by the composer Frédéric Chopin. Published between 1830 and 1880, these scores provide an important resource for studying Chopin's publishing history, and include many examples of works published concurrently in different countries with textual variations. To make this collection more accessible to scholars, the collection is being digitized and made available over the Web.

2 Building the online collection

The Preservation Department at the University of Chicago Library is scanning these scores according to guidelines published by the National Archives and Records Administration (1998). Scores are scanned at a resolution of 400dpi, which the department determined would capture the smallest significant details. Digital images are stored as TIFF files in 24-bit RGB with no compression. The TIFF images are not edited by hand; flawed images are rescanned rather than retouched. For web delivery, two JPEG derivatives are made for each TIFF image, one 2000 pixels wide and the other 700 pixels. The larger image shows greater detail, but the smaller image is faster to download and easier to view on common monitors. Table 1 shows average and extreme image file sizes. The smallest JPEG files are for blank pages, the largest are for pages of fine text. As of Aug. 12, 2003, 370 scores have been scanned, and a total of 7031 TIFF images created, for an average of approximately 19 scans per score. A total of 13,030 derivatives have been produced, as producing derivatives lags behind scanning. Testing has begun to add derivatives in the DjVu format, which provides a variety of desirable client-side features, such as page-turning, zooming, and scale-to-fit printing (Bottou, et al., 1998).

Format	Avg. size	Min. size	Max. size
TIFF 400dpi	82.4MB	56.1MB	96.8MB
JPEG 2000px	411KB	272KB	1896KB
JPEG 700px	70KB	40KB	344KB

Table 1: Image file sizes

Detailed bibliographic records in the UC library catalog were created or enhanced by a professional music cataloger. These MARC records provided the descriptive metadata for the online collection. The records strictly follow AACR2, in accordance with guidelines for submission to OCLC. Uniform titles were created as needed, and are used to gather different editions of a score together.

A computer program assembles the digital scores as XML documents conforming to the Metadata Encoding and Transmission Standard (METS). Structural and administrative metadata for a score are recorded in a relational database at the time of scanning. These data are exported and used by the program to populate the METS document for the digital version of the score. The bibliographic metadata in MARC are translated into the Metadata Object Description Schema (MODS) and embedded in the METS document. The completed METS documents are the fundamental representations of the digital scores for this project and can be repurposed as appropriate.

The user interface to the collection is provided by Greenstone (Witten, Bainbridge & Boddie, 2001). The METS digital score objects are transformed via XSLT into the Greenstone native document format. The flexibility inherent in Greenstone permits construction of custom intra-document navigation mechanisms specific to this collection. Metadata in the documents programmatically massaged for improved

retrieval. For example, a place name thesaurus is simulated, so a search for scores published in Leipzig will match the six variations of that city name found in the database.

3 Usage analysis

Analysis of web server logs for the first 5 months of public use yield some interesting and thought-provoking findings. Unless otherwise indicated, all numbers exclude on-campus use. The site averages about 100 hits (collection browsing or score viewing operations) daily. Connections have been made from over 900 individual IP addresses. At least 30% of total use appears to be international.

Table 2 shows the proportion of user time spent navigating and viewing scores, broken down by the search interface and the four browse listings. The user interface permits keyword searching of 11 indexes; browsing by Title, Uniform Title, Genre, or Dedicatee; and viewing the score images. The Title and Genre browse listings are the most-used ways of accessing the collection. This is consistent with the informal user testing conducted with music students at the bachelor and doctoral levels. The View/Navigation ratio for the Genre browse listing suggests it is the most effective route by which users access scores that they will examine in more detail.

During score-viewing activity, over 17,000 score images were downloaded. 79% of images downloaded were the 700 pixel wide JPEG files and 21% were the 2000 pixel wide JPEG files. Currently, the smaller image is displayed by default and the user has to take an extra action to see the larger image. These numbers suggest that users are very interested in the larger images and that the interface should make them more readily accessible.

Table 3 shows the frequency with which each keyword index has been searched within the search interface. Aside from Title, which is the most prominent index, the frequencies bear no relation to the order of indexes presented on the search page. The popularity of the Opus keyword search is consistent with user requests to improve access by opus number in other parts of the interface.

Access method	Viewing scores	Navigating collection	View/Nav. ratio
Title	5592	1944	2.88
Genre	2540	646	3.93
Keyword Search	771	648	1.19
Uniform Title	996	639	1.56
Dedicatee	861	285	3.02

Table 2: User activity by access method (search interface and browse listings)

4 Future directions

The user interface to this collection will be refined based on further user studies, user feedback, and information collected from web logs. User studies and feedback have already suggested changes that are planned for a future revision. For example, the organization of some of the browse listings, and the mechanisms for seeing different versions of the score images will change.

The descriptive metadata for the scores in this collection will be made available for OAI harvesting.

We are exploring options to add content-based search features to the collection. We would like to provide the searching of the musical symbolic content of the scores. Options being explored include use of optical music recognition to create searchable representations and/or acquisition of pre-existing symbolic representations (i.e., MIDI, GUIDO, etc.). Adding audio files to complement the scores is also under consideration.

This may also be an appropriate test collection for various techniques in music information retrieval. The varied music typography and complexity of notation may be especially useful for testing progress in optical music recognition. The presence of scores with multiple variant texts and full feature bibliographic data provides opportunities for gathering and distinguishing between variations of the same work.

The Chopin Early Editions collection is available at `http://chopin.lib.uchicago.edu/`.

Keyword Index	Percentage	Occurrences
Title	54%	348
Opus No.	12%	80
General Keyword	10%	64
Date	5%	33
Genre	5%	31
Subject	5%	30
Uniform Title	2%	16
Publisher	2%	13
Place of Publication	2%	12
Plate Number	2%	12
Dedicatee	1%	9

Table 3: Searches by keyword index

References

Bottou, L., Haffner P., Howard P., Simard P., Bengio, Y. & LeCun, Y. (1998). High quality document image compression with `DjVu'. *Journal of Electronic Imaging.* 7(3), 410-425.

Metadata Encoding and Transmission Standard (METS). Library of Congress. Retrieved Aug 11, 2003 from http://www.loc.gov/standards/mets/

Metadata Object Description Schema (MODS). Library of Congress. Retrieved Aug 11, 2003 from http://www.loc.gov/standards/mods/

National Archives and Records Administration. (1998). *NARA Guidelines for Digitizing Archival Materials for Electronic Access*. Retrieved August 10, 2003 from `http://www.archives.gov/research_room/arc/arc_info/guidelines_for_digitizing_archival_materials.pdf`

Witten, I. H., Bainbridge, D. & Boddie, S. J. (2001). Greenstone: Open-Source Digital Library Software. D-Lib Magazine. 7(10). Retrieved August 10, 2003 from `http://www.dlib.org/dlib/october01/witten/10witten.html`