SEARCHING THE LIBER USUALIS: USING COUCHDB AND ELASTICSEARCH TO QUERY GRAPHICAL MUSIC DOCUMENTS

Jessica Thompson
BMARS
Department of Music
Dartmouth College
jessica.thompson.gr@dartmouth.edu

Andrew Hankinson
CIRMMT
Schulich School of Music
McGill University
andrew.hankinson@music.mcgill.ca

Ichiro Fujinaga
CIRMMT
Schulich School of Music
McGill University
ich@music.mcgill.ca

1. INTRODUCTION

Our long-term goal is to make the text and melodies of all digitized graphical music documents available online and searchable. This demonstration represents our first step towards this goal: a fully searchable version of the 1961 edition of the *Liber Usualis*, a 2340-page book of Gregorian chant, available via a web application[1]. After digitized page images have been automatically transcribed using optical music recognition (OMR) and optical character recognition (OCR) software, we use CouchDB and ElasticSearch to store and query our data. Search results are highlighted on the original page images (Figure 1).

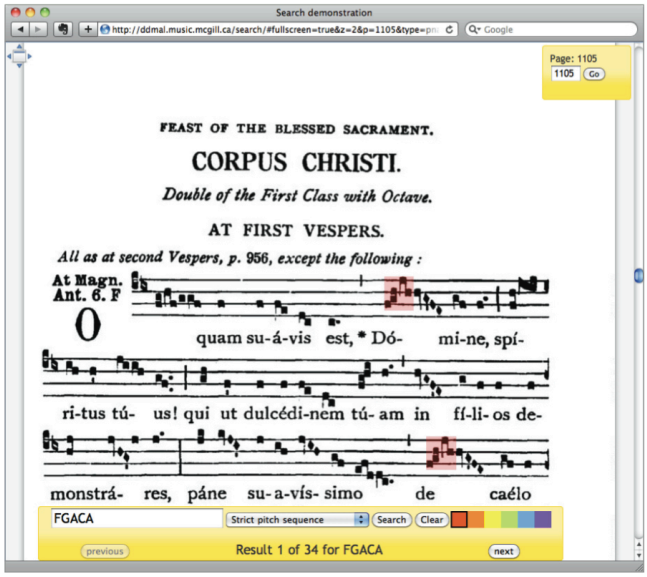

Figure 1. Screen shot of the *Liber Usualis* web application. Instances of the melody "FGACA" are highlighted on the original page image.

[1] Demo can be found at:
http://ddmal.music.mcgill.ca/liber/

2. PREVIOUS WORK

The Global Chant Database (GCD) (http://www.gregorian-chant.org/) is a valuable online resource for musicologists who wish to find plainchant melodies in medieval sources and modern editions. The search interface gives the user access to almost 25,000 "incipits," or the beginnings phrase of a chant, but does not provide access to the full musical content. The GCD offers a version of the *Liber Usualis* that is presented as individual image files. However, manual transcription of the musical contents of the entire *Liber* would be extremely labour intensive, and manual transcription of all musical sources for the purposes of retrieval would be a monumental, if not impossible task.

3. WORKFLOW

3.1 Overview

To address the need for an automated method of printed music transcription, we devised a workflow (Figure 2) that uses existing tools to transform a set of document images into a fully searchable representation of the Liber Usualis. This book was chosen because of its importance as a reference book, its use of monophonic neume notation made the transcription task easier, and the mixture of text and music made it a good candidate to test both OMR and OCR technologies. One key feature of our system was that we wished to search the content of the images *in situ*—that is, store both the musical and textual content and their locations on the page image. For this, we needed to be able to index and quickly retrieve all pitch and text sequences and their locations on every page.

3.2 Chant and Text Recognition

Page layout analysis and segmentation was performed automatically using Aruspix (http://www.aruspix.net/) and then manually corrected to ensure all of the different page components (music, lyrics, text, etc.) were correctly identified. These elements were extracted as separate layers. The music layer was sent to Gamera (http://gamera.informatik.hsnr.de/) for automatic recognition of the neume shapes. Any misrecognized shapes were corrected manually. We developed an

Figure 2. Overview of workflow from digitized image files to searchable content via a web application.

algorithm to recognize the pitch of each neume shape automatically [1]. The recognized neume shapes and pitches were fed back into Aruspix for manual verification and correction. Aruspix saved the final output, including co-ordinate information for every page element, in MEI [2] files. For the text layers we used Ocropus, an open-source tool for optical character recognition, to automatically extract and recognize the text. No manual correction was performed on the text, but the recognized words were auto-corrected against a Latin dictionary.

3.3 CouchDB

Our CouchDB databases were designed for fast and easy retrieval of melodic and textual information. We used an *n*-gram approach in which a CouchDB record, or "document," was stored for every pitch sequence the user might want to retrieve (Figure 3). CouchDB documents were divided into separate databases based on pitch sequence length—one database for 2-grams, one for 3-grams, up to 10-grams. As a result, only one database needs to be consulted for a search string of a known length, minimizing the number of documents to be considered. This is a storage-heavy solution but it results in very fast retrieval. This indexing resulted in a database containing 2 945 895 individual note-gram documents for all nine document databases.


```
{
   "_id": "872a10d7d4dae30795c40afb2d97eb46",
   "_rev": "2-e0a5c321330064f0ff7114c41830de95",
   "pagen": 1105,
   "semitones": "2_2_3_-3",
   "pnames": "fgaca",
   "neumes": "scandicus_clivis",
   "intervals": "u2_u2_u3_d",
   "location": [
       {
           "width": 79,
           "ulx": 1153,
           "uly": 1153,
           "height": 88
       }
   ],
   "contour": "uuud"
}
```


Figure 3. A sample 5-gram document stored using CouchDB and indexed using ElasticSearch.

3.4 ElasticSearch

ElasticSearch is an open source, distributed, RESTful, search engine built on top of Lucene (http://lucene.apache.org/). ElasticSearch's CouchDB River feature automatically indexes CouchDB using the `_changes` stream provided by CouchDB. The River feature allows us to have continuous synchronization between our CouchDB databases and our ElasticSearch indices.

3.5 Web Application

Our web application uses a high-resolution, multi-page document viewer, Diva (Document Image Viewer with Ajax) [3], to display page images. Search melodies are specified with character strings such as "FGACA." When a query is sent to ElasticSearch, it returns the page coordinates of the results so that they can be highlighted on the original image file.

4. CONCLUSION

To the best of our knowledge, this is the first system to use CouchDB and ElasticSearch to store and search symbolic music data and the first web application to provide searchable images of plainchant notation. Our automated tools distinguish this project from other similar projects and will ultimately allow us to achieve our goal of making the text and melodies of all digitized music available online and searchable.

5. ACKNOWLEDGEMENTS

We would like to thank our outstanding development team for their hard work: Remi Chiu, Mahtab Ghamsari, Jamie Klassen, Saining Li, Wendy Liu, Mikaela Miller, Laura Osterlund, Alastair Porter, Laurent Pugin and Caylin Smith. We would also like to thank the Social Sciences and Humanities Research Council of Canada for funding this project.

6. REFERENCES

[1] G. Vigliensoni, J. A. Burgoyne, A. Hankinson and I. Fujinaga: "Automatic Pitch Recognition in Printed Square-Note Notation," *Proc. ISMIR*, forthcoming, 2011.

[2] P. Roland: "The Music Encoding Initiative (MEI)," *Proc. First Int'l Conf on Musical Applications Using XML,* pp. 55–59, 2002.

[3] A. Hankinson, W. Liu, L. Pugin, and I. Fujinaga: "Diva.js: A Continuous Document Viewing Interface," *Code4lib Journal*, No. 14, 2011. http://journal.code4lib.org/articles/5418