DIGITAL DOCUMENT IMAGE RETRIEVAL USING OPTICAL MUSIC RECOGNITION

Andrew Hankinson **John Ashley Burgoyne** **Gabriel Vigliensoni**
Alastair Porter **Jessica Thompson** **Wendy Liu**
Remi Chiu **Ichiro Fujinaga**

Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT)
Schulich School of Music, McGill University, Montréal, QC, Canada
andrew.hankinson@mail.mcgill.ca

ABSTRACT

Optical music recognition (OMR) and optical character recognition (OCR) have traditionally been used for document transcription—that is, extracting text or symbolic music from page images for use in an editor while discarding all spatial relationships between the transcribed notation and the original image. In this paper we discuss how OCR has shifted fundamentally from a transcription tool to an indexing tool for document image collections resulting from large digitization efforts. OMR tools and procedures, in contrast, are still focused on small-scale modes of operation. We argue that a shift in OMR development towards document image indexing would present new opportunities for searching, browsing, and analyzing large musical document collections. We present a prototype system we built to evaluate the tools and to develop practices needed to process print and manuscript sources.

1. INTRODUCTION

Optical character recognition (OCR) is used to convert digital images of text into computer-manipulable representations, which in turn are used to store and index the content of books, newspapers, scholarly journals, and magazines. OCR has been integrated into many large-scale print digitization initiatives, and is currently being used to provide users with the unprecedented ability to search and retrieve millions of sources instantly—a task that previously would have taken many lifetimes.

While OCR is opening up new avenues for users to search, discover, and analyse large quantities of textual material, the content of printed music documents is still trapped almost entirely in the physical world. Even collections that have been digitized and placed online are still merely pictures of pages, with no means of extracting their contents. Unfortunately, current optical music recognition (OMR) software packages have not been designed to process large volumes of page images efficiently. Rather, they are still designed for small-scale, single user transcription. In order to provide OMR tools comparable to the OCR tools now available there is a need for developing tools, technologies, and best practices for recognizing, indexing, searching, and retrieving large amounts of digital page images.

We believe that large-scale OMR projects are critical to research in MIR and also computational musicology. The vast majority of human musical output from the past 1000 years does not exist in any kind of manipulable digital format but rather lies within the enormous collections of printed music and music manuscripts sitting on library shelves across the globe. Relying on humans to transcribe and share this music is expensive and unsatisfactory for many purposes. Many human transcriptions, especially of early music, involve a substantial amount of personal interpretation on the part of the transcriber, and end users may not entirely agree with a particular interpretation. Moreover, many musicologists are interested in studying how the extra-musical content on a page informs the musical content, and these scholars need to be able to access the orginal page images in order to draw concusion. By combining automatic transcription with retrieval of the original page image, we can build systems that permit users to retrieve documents by content but then rely on the original image for study, relaxing the need for exact or "objective" transcription. Analogous to the situation with OCR, even lower-accuracy OMR would be sufficient to direct a user toward documents of interest, an order of magnitude more quickly than the current situation whereby researchers must visit library shelves and manually study every page.

In this paper we argue that there are a number of OMR technologies that must be in place to enable mass music recognition and retrieval projects. We will discuss the development of similar OCR technologies built to support large-scale text document image indexing and retrieval systems and compare that to the existing OMR technologies. As part of this discussion, we present a prototype project developed as part of the Single Interface for Music Score Searching and Analysis (SIMSSA) initiative. This project includes the development of an

OMR workflow system, a notation encoding format for storing the results of our OMR system, and the development of search and retrieval tools. We discuss our findings from this prototype, and present some of opportunities for future work.

2. PREVIOUS WORK

Both OCR and OMR were initially conceived as transcription technologies; that is, a digital page image was supplied, the textual or musical content was extracted, and then the original image was discarded. The result was the transcribed content of the page in a format suitable for further editing in a word processor or notation editor. This method left no direct correspondence between the page content and its location on the original image in the output format.

When computing advanced enough to display images, OCR began to be used as a means of navigating and retrieving document images based on their textual content. Although the recognition process remained the same, OCR file formats began preserving correspondence between document content—words, paragraphs, columns, graphics—and the original page image. When combined with a search engine, this provided users with the ability to navigate to the exact occurence of a search word or phrase. This was the emergence of the shift in OCR from a technology that "merely" transcribed text, to one that permitted navigation through large numbers of digital document images. Many researchers at this conference have probably experienced the benefits of these systems for locating journal articles or conferences papers in systems like JSTOR.

In the next section we will look at a number of technologies in OCR that have supported its emergence as a document image navigation technology. We will then consider a few of the projects that have tried to do similar things with printed music documents.

2.1. Document Image Formats

One of the most crucial components of a document image indexing system is the ability to correlate recognized objects with their location on the page. In this section we discuss a few formats developed for textual document indexing.

Nagy [1] describes an early system for using OCR on document images from pages of technical journals, allowing users to search and retrieve image segments containing their query terms. He notes that a "major strength of the approach is the preservation of the original layout of the documents, which not only augments reading comprehension but also often conveys indispensable information on its own." This seems to be the first such mention of retrieving document page images using OCR analysis.

Stehno et al. [2] describe the METS/ALTO format (Metadata Encoding and Transmission Standard / Analyzed Layout and Text) for mapping layout structures and text passages from book pages. This is currently the most widely used standard for preserving text layout derived from OCR systems.

Several other formats deserve a brief mention as well. The hOCR format is supported by the OCRopus and Tesseract software, used by the Google Book digitization project [3]. Portions of the Internet Archive digitization project use the DjVuXML format which contains the words and coordinates of a "hidden" layer [4]. The HathiTrust Project uses their own XML format.

The PAGE format [5] is designed to encode data for evaluating OCR document analysis. It is different than most other formats presented here, in that it encodes ground-truth data for evaluating document layout. This includes encoding features like reading order (the order in which columns or segments of a page are read by a human).

In all of these cases, document layout and text content are maintained in an integrated document format, allowing the OCR system to store image coordinates for each document element (words, lines, paragraphs, etc.) recognized by the analysis software. In the next section we will briefly discuss how this data may be used in an indexing system to retrieve page images.

2.2. Document Image Retrieval

Indexing for page image retrieval is more complex than indexing for simple text retrieval. An index must be built containing all the words that have been recognized from the images, but these must be further correlated with their page and location on the page image. Retrieving page images requires indexing and storing key words and their positions in the document images where they occur. This gives users the ability to enter a search query (a word or phrase) and retrieve the pages where the result of their query can be found, highlighting their exact positions on the page image.

The HathiTrust has constructed a correlated text and image index for their collection. They incorporate the graphical locations for each recognized word into their index [6]. They note that the coordinate positions for every word accounts for 85% of the index size of a particular book. This means that for their required goal of 10 million books, their expected index size is two terabytes of which most of the information is OCR coordinate data.

2.3. Large-Scale OCR

To handle large numbers of documents, OCR applications have moved away from standalone desktop applications to server-based solutions. This allows distributed task separation, whereby multiple teams can simultaneously work on digitization, recognition, correction, and publishing without being bound to a single workstation. Many tasks can be partially or fully automated, requiring human intervention only as a quality-control measure.

In the commercial sector there are a number of large-scale solutions. Perhaps the most successful example of

this is the Abbyy Recognition Server[1], a centralized OCR workflow management system. Documents are ingested by digitization, automatically recognized, then verified, corrected, and further indexed by humans sitting at multiple workstations. For open-source software, there are a number of command-line tools that can be chained together to form an automated OCR system. The OCRopus and Tesseract systems [7], developed by Google for their book search projects, contains a number of tools for creating highly customizable OCR systems. The recently-completed IMPACT (Improving Access to Text) project [8], a €16.5M research project, focused on building new OCR tools and best practices for libraries and archives. They have created a system that allows multiple image processing, OCR, and results evaluation tools to be chained together to form an *ad hoc* recognition system.

Since recognition systems will never be perfect, techniques that enable humans to correct OCR and ensure that recognition errors will not create problems for retrieval. Correction, however, can be very time and labour intensive. Some unique solutions have been developed to help offset the costs of this task. The Australian Newspaper Project [9] has created a "crowdsourced" correction system, where more than 9,000 volunteers have now corrected more than 12.5 million lines of text, with more corrections added all the time. The reCAPTCHA project [10] has produced over 5 billion human-corrected OCR words by presenting the correction task as a spam-fighting challenge to prove that the corrector is a human and not an automated system. Tools for distributed proofreading and correction allow for a constantly-improving search and retrieval system, and also for the collection of pixel-aligned ground-truth data to further improve the accuracy of OCR systems.

2.4. Music Document Image Retrieval

The purpose of providing a review of tools and techniques employed by text search projects is to compare and contrast it to similar work done for OMR. Unlike the large text initiatives, such as the HathiTrust, Google Books, and Internet Archive projects, we are unaware of any publicly available databases that allow users to retrieve page images from printed books based on automatic transcription of the page contents using OMR. There are, however, a few projects that have developed some functionality worth mentioning here.

The PROBADO Music Project [11],[12],[13] [14] is perhaps the largest and longest-running project incorporating large-scale OMR for use in search systems. This project seeks to provide a unified interface for retrieving symbolic and audio representations of music pieces. As of October 2010, their dataset consisted of 50,000 pages from 292 books. The content of their dataset is music printed in common Western notation in a variety of genres and instrumentations, including opera, symphonic works, and Classical and Romantic piano music.

The primary goal of the PROBADO project is to allow symbolic, image, and audio synchronization, providing users with the ability to navigate a score and hear the audio, or navigate the audio and jump to its corresponding position in the score. Their technique generates MIDI files from OMR, rendered to an audio representation, and then aligned with different audio recordings of the work. The audio is then aligned at the measure level with a score image, allowing the system to highlight the current measure as the audio plays.

The PROBADO project uses the SharpEye ASCII file format for storing the notation-to-pixel coordinate information. This format is documented at [15], but is only supported by the SharpEye OMR system. Similarly, Hankinson et al. [16] propose the use of the Music Encoding Initiative (MEI) format for maintaining notation-to-pixel correspondence.

Bainbridge et al. [17] describe a Greenstone plug-in utilizing the CANTOR OMR system. Their system transcribes the notated music and makes it available for searching. In their system they make the original page image available for viewing. Unfortunately, development on this system seems to have stopped, and no working version of their retrieval system can be found.

3. LU PROTOTYPE

We have created a prototype system providing notation-based retrieval of document images in a web application. The *Liber Usualis* (LU) [18] is a liturgical service book produced by the Roman Catholic church and an important source for Gregorian chant. It uses square-note neume notation derived from the earlier Franconian style but modernized by the monks at Solesmes, France in the late 19th Century. There has been very little work on OMR for this type of notation, with the exception of [19]. We performed OMR and OCR on all 2,340 page images in this book, maintaining notation and image correspondence. We then developed a web application that allows basic query input based on *n*-gram indexing of the notation content, highlighting the locations of results *in situ* on the page image. In this section we will briefly review the components of this prototype. A full overview may be found in [20], and some details of the specific technologies we developed may be found in [21], [22]. We have made a public demo of our retrieval system available online.[2]

3.1. OMR Workflow

Our workflow is illustrated in Figure 1. We begin with a page image, captured by either scanning or photographing a book. In the case of the LU, we began with a complete PDF file of pre-scanned images. We then sent each page through the workflow.

[1] http://www.abbyy.com/recognition_server/

[2] http://ddmal.music.mcgill.ca/liber

The first step was automatic page layout analysis, which we used to separate the textual and musical areas of the page. We performed the layout analysis using a version of Aruspix, an application originally designed for OMR of Renaissance printed music, which we modified to operate on the neume notation in the LU. The page images were automatically scaled and straightened. Aruspix is capable of automatically locating and identifying various graphical page elements, providing the ability to distinguish between musical and textual content: musical staves, lyrics, ornate letters, lyrics, title elements, or other text. The different page elements are given different pixel colours after the automated analysis, creating separable "layers" that contain either exclusively music or exclusively text elements. The automated analysis saves a considerable amount of time and labour, although any mis-classified page elements do need to be corrected manually. Figure 2 shows the correction interface in Aruspix, with a pop-up context menu allowing the operator to select an area of the image and re-classify it as a different type of page. The LU does not have a particularly complex layout, and most pages took between 30 and 130 seconds to correct (median 77 s).

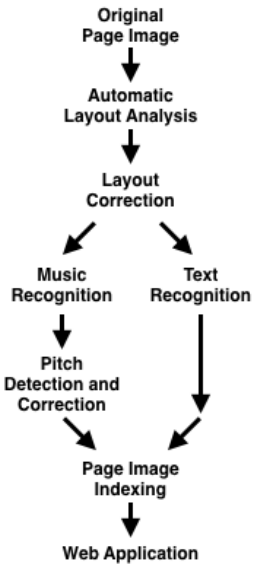

Figure 1: OMR Workflow for the Liber Usualis

Following the layout analysis, the text layers were sent through an OCR workflow stage, which allowed the text to be searched and linked to the specific regions of each page on which the text occurred. We used OCRopus, a third-party open-source OCR engine, to perform the text analysis. Minimal work was done to correct the OCR output, however. Post-OCR, we used a simple edit distance to auto-correct recognized text from a dictionary of liturgical Latin words. Lyrics that were broken into syllables were automatically re-joined at a hyphenated break to form complete words. No further human correction or processing was performed. The resulting text has a large number of errors, but it was sufficient for a "proof of concept." The output of this stage was fully OCRed text lines with the bounding-box co-ordinates for the full line.

The music layer was sent through an OMR workflow. Using the Gamera toolkit [23], we first removed the staff lines from each musical layer. Removing staff lines not only facilitated OMR but also allowed us to compute precise bounding boxes for each musical element. These bounding boxes are essential information for retrieval systems that wish to show the results of musical queries on the page. Gamera uses adaptive *k*-nearest-neighbour classifiers for musical symbols, improving its classification performance by using the information from the errors corrected on previous pages. It took between 7 and 16 minutes to correct the errors on most pages (median 11 min). Gamera keeps track of the location of staff lines when it removes them, but it classifies on the shapes, not the pitches of musical symbols. The last step of the OMR workflow was a customized algorithm for combining information about the location of the staff lines, the bounding box for each shape, and the shape of the musical symbol itself to add pitch information to every symbol.

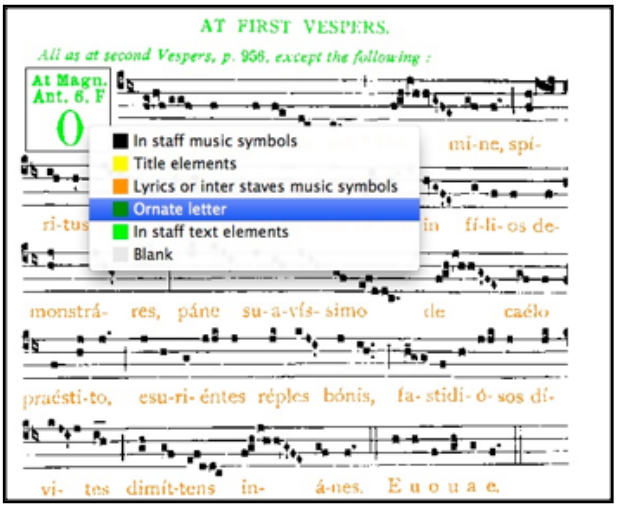

Figure 2: Layout Analysis and Correction in Aruspix

After extracting the pitch information and bounding boxes for every musical symbol and text line on every page, the final step of the workflow was to store the recognized page content into a standard file format. We chose the MEI format for a number of reasons. MEI is an XML-based notation encoding scheme, but unlike most notation formats it can be extended to support many different types of notation [24]. This was particularly valuable in our case, since the neume notation used by the LU is a revival of much older plainchant notation with additional symbols added to indicate breathing marks or articulations. With MEI we were able to develop a custom encoding scheme to support neume notation markup, while maintaining the broader document framework and markup structure of MEI.

To support document image indexing and retrieval, MEI provides the ability to define image zones—pixel-based bounding boxes that store co-ordinates on a reference image—and correlate them with the recognized

musical and textual elements. It is important to note here that this is done while still preserving the musical structure; that is, the notation maintains the melodic and symbolic structures that are expected from a notation encoding scheme. Each musical and textual element is then correlated with a defined zone using the MEI `@facs` attribute. The end result is an XML hierarchy containing both the musical, textual, and graphical information correlated and ready to be indexed by a search engine to facilitate image retrieval.

3.2. Indexing, Searching and Retrieval

One of the most important musicological uses for the LU is as a compendium of important chant melodies that appear in some form across a wide variety of ancient musical manuscripts and many later compositions. Despite its importance, there is no thematic catalogue of its musical content, and at more than 2,000 pages, it can be very time-consuming for researchers or musicians to find what they are looking for. To facilitate retrieval we created an efficient index for retrieving musical fragments from across the LU while maintaining information the location of these melodic fragments on their respective pages. Following Stephen Downie [25], we generated indexes on *n*-grams, for *n* from 2 to 10, on the following five features:

- pitches, a concatenated string of all of the pitch names;
- intervals, represented as the directed melodic intervals between successive pitches in musical steps;
- semitones, represented as the directed melodic intervals between successive pitches in semitones;
- contour, represented as sequences of "up," "down," or "repeated," i.e., the direction of the intervals; and
- neumes, represented by their component neume names, i.e., if an *n*-gram was represented by the neume sequence "punctum clivis clivis."

We also indexed the textual content of each page, including the co-ordinate information for each recognized line.

For each *n*-gram, the index also included the page on which an item appeared and its aggregate bounding box. Combined, the indexes include approximately three million unique *n*-grams, which we store in an Apache Solr instance[3].

Users do not interact directly with the Solr instance. We built a web application based on the open-source Diva.js viewer [26] to present the original document images and highlight the results of queries on these images. The web application uses the indexed *n*-grams to provide a number of search capabilities:

- strict or pitch-invariant sequences, where the user can type in a sequence of pitch names and it will either search for the literal string of pitches, or use the semitone index to search for possible matches that use the same intervallic content but contain different pitches;
- contour, containing the rough shape of the target melody, e.g., "dduurr";
- intervals, containing the specific shape of the target melody, e.g., "d2 d2 u2 u3 r r";
- neumes, where the user specifies a sequence of neume names, e.g., "punctum clivis clivis"; and
- text, for retrieving pages based on lyrical or textual content.

Using the co-ordinate data from the OMR and stored in the MEI, the results from a users' query for a pitch sequence will bring the user to the page where their result can be found, with the bounding box around the search result highlighting their query. Figure 3 shows a screenshot from our web application with the result of the pitch-sequence query "*edcdeee*" highlighted on the original page image of the LU.

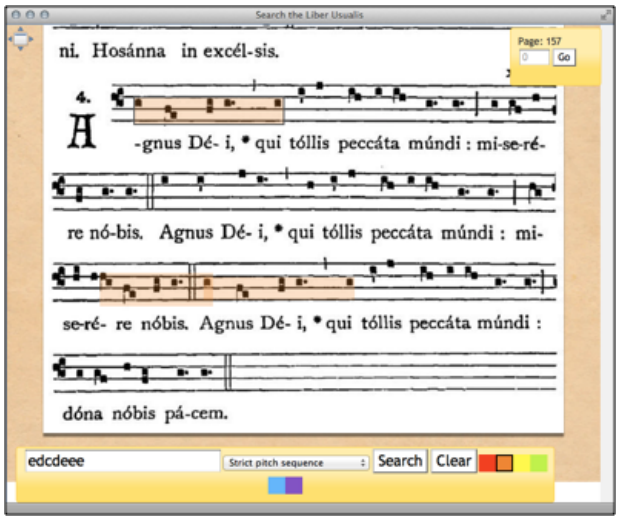

Figure 3: The Liber Usualis Interface

4. DISCUSSION AND CONCLUSION

The SIMSSA initiative is a long-term research program for investigating and supporting large-scale OMR and document image retrieval for all types of music documents, from early manuscripts through to modern music notation. The work presented in this paper is an initial attempt at building systems that support processing and retrieval at a scale and quality level that, to date, has not been achieved for musical documents. In this paper we have identified technologies and techniques that have been developed to support OCR for transcribing and navigating large numbers of document images, and have demonstrated a prototype system we have developed as a platform for further research into how to shift OMR from small-scale transcription to large-scale document image navigation and retrieval.

There are many open research questions arising for this work that need further investigation. One of the most critical is the need for user studies and experimental interfaces for musical document retrieval. Most current symbolic search systems are built around query interfaces that provide limited access to the underlying

[3] http://lucene.apache.org/solr/

notated music, typically restricted simple pitch or rhythm queries. More robust systems must be built to support more complex analysis-retrieval tasks, such as investigating the occurrence of specific cadential patterns or movement between multiple simultaneous voices. More complex query and analysis systems will in turn require more sophisticated user interfaces, which will need a deeper understanding of what what types of questions musicologists, theorists, and performers would like to see supported in a retrieval system.

In our opinion, OMR must move beyond desktop applications and simple transcription. New modes of operation can and should be developed, including server- and browser-based recognition, distributed proofreading and correction, networked recognition systems, and expanded research on recognition evaluation by building curated ground-truth datasets covering different styles and types of music notation. Our prototype system represents a first step towards investigating many of these topics, and through the SIMSSA project we hope to spur further research to make the world's music collections available to all.

5. ACKNOWLEDGEMENTS

This work would not have been possible without the efforts of a number of people. The authors would like to thank Julie Cumming, Mahtab Ghamsari, Jamie Klassen, Saining Li, Mikaela Miller, Laura Osterlund, Laurent Pugin, and Caylin Smith for their contributions. The SIMSSA project is funded by the Social Sciences and Humanities Research Council of Canada. Further funding was provided by the Centre for Interdisciplinary Research in Music Media and Technology and the Canadian Foundation for Innovation.

6. REFERENCES

[1] Nagy, G. 1992. *Towards a structured-document-image utility*. In *Structured Document Image Analysis*, eds H. Baird, H. Bunke, and K. Yamamoto. Berlin: Springer.

[2] Stehno, B., A. Egger, and G. Retti. 2003. METAe—Automated encoding of digitized texts. *Literary and Linguistic Computing* 18 (1): 77–88.

[3] Breuel, T., and U. Kaiserslautern. 2007. The hOCR microformat for OCR workflow and results. In *Proc. Int'l. Conf. on Document Analysis and Recognition*, 1063–7.

[4] Kumar, R. 2008. Bulk access to OCR for 1 million books. http://blog.openlibrary.org/2008/11/24/bulk-access-to-ocr-for-1-million-books/.

[5] Pletschacher, S., and A. Antonacopoulos. 2010. The PAGE format framework. In *Proc. Int'l. Conf. on Pattern Recognition,* Istanbul, TR. 257–60.

[6] Farber, P. 2009. Large-scale full-text indexing with Solr. http://www.hathitrust.org/blogs/large-scale-search/large-scale-full-text-indexing-solr.

[7] Breuel, T.. 2009. Recent progress on the OCRopus OCR system. In *Int'l. Workshop on Multilingual OCR*, Barcelona, ES. 1–10.

[8] Balk, H., and L. Ploeger. 2009. Impact: Working together to address the challenges involving mass digitization of historical printed text. *OCLC Systems and Services* 25 (4): 233–48.

[9] Holley, R. 2009. Many hands make light work: Public collaborative OCR text correction in Australian historic newspapers. *National Library of Australia Staff Papers*.

[10] Von Ahn, L., B. Maurer, C. Mcmillen, D. Abraham, and M. Blum. 2008. reCaptcha: Human-based character recognition via web security measures. *Science* 321 (5895): 1465–8.

[11] Diet, J., and F Kurth. 2007. The PROBADO music repository at the bavarian state library. In *Proc. Int'l. Conf. Music Information Retrieval,* Vienna, AT.

[12] Kurth, F., M. Müller, C. Fremerey, Y. Chang, and M. Clausen. 2007. Automated synchronization of scanned sheet music with audio recordings. In *Proc. Int'l. Conf. on Music Information Retrieval* , Vienna, AT. 261–6.

[13] Fremerey, C. 2010. Automatic organization of digital music documents: Sheet music and audio. PhD diss., Mathematics and Natural Sciences, University of Bonn, Bonn, DE.

[14] Damm, D., F. Kurth, C. Fremerey, and M. Clausen. 2009. A concept for using combined multimodal queries in digital music libraries. *Research and Advanced Technology for Digital Libraries* 261-72.

[15] Jones, G. OMR engine output file format. http://www.visiv.co.uk/tech-mro.htm.

[16] Hankinson, A., L. Pugin, and I. Fujinaga. 2010. An interchange format for optical music recognition applications. In *Proc. Int'l. Society for Music Information Retrieval*, Utrecht, NL.

[17] Bainbridge, D., C.G. Nevill-Manning, I.H. Witten, L.A. Smith, and R.J. Mcnab. 1999. Towards a digital library of popular music. In *Proc. ACM Conf. on Digital Libraries*, Berkeley, CA. 161–9.

[18] Catholic Church. 1963. *The Liber Usualis, with introduction and rubrics in English*. Tournai, Belgium: Desclée.

[19] Ramirez, C., and J. Ohya. 2011. OMR of early plainchant manuscripts in square notation: A two-stage system. In *Proc. SPIE*, 1-10.

[20] Hankinson, A., J.A. Burgoyne, G. Vigliensoni, and I. Fujinaga. 2012. Creating a large-scale searchable digital collection from printed music materials. In *Proc. Advances in Music Information Research*, Lyon, FR.

[21] Hankinson, A., P. Roland, and I. Fujinaga. 2011. The music encoding initiative as a document encoding framework. In *Proc. Int'l. Society for Music Information Retrieval*, Miami, FL.

[22] Vigliensoni, G., J.A. Burgoyne, A. Hankinson, and I. Fujinaga. 2011. Automatic pitch detection in printed square notation. In *Proc. Int'l. Society for Music Information Retrieval*, Miami, FL.

[23] Macmillan, K, M Droettboom, and I Fujinaga. 2001. Gamera: A structured document recognition application development environment. In *Proc. Int'l. Symposium on Music Information Retrieval*, Bloomington, IA. 15–6.

[24] Hankinson, A., P. Roland, and I. Fujinaga. 2011. The Music Encoding Initiative as a document encoding framework. In *Proc. Int'l. Conf. Music Information Retrieval*, Miami, FL.

[25] Downie, S. Evaluating a simple approach to music information retrieval: Conceiving melodic *n*-grams as text. PhD diss., University of Western Ontario, London, Ontario.

[26] Hankinson, A., W. Liu, L. Pugin, and I. Fujinaga. 2011. Diva.Js: A continuous document image viewing interface. *Code4lib Journal* 14.