Beyond transcription: Case studies in special document analysis requirements

Michael Droettboom
Digital Knowledge Center, Sheridan Libraries
The Johns Hopkins University
3400 N. Charles St.
Baltimore, MD 21218

Abstract

Through case studies of two current projects at Johns Hopkins University's Digital Knowledge Center (DKC), this paper discusses some document analysis applications other than straightforward transcription that were developed through direct communication with users of digital collections. These applications include lyrics extraction from sheet music, an image-based annotation collaboratory environment and automatic illumination-finding in medieval manuscripts. Some of the problems encountered developing new technology across a disciplinary divide are then discussed. This paper aims to foster discussion related to how to make document image analysis research more user-centric.

1. Introduction

When creating digital libraries of cultural heritage materials, the purpose of document image analysis is often to create an "equivalent" symbolic transcription in machine-readable form. Of course, there are always many complications and trade-offs associated with defining exactly what an "equivalent" representation is. However, ignoring that issue, this kind of transcription, even if fairly lossless, can be less important than other kinds of extractable information. There are often other more specialized uses of document image analysis that are both more relevant to users of the documents and more technically tractable.

This paper will discuss research involving two very different collections, the Lester S. Levy Collection of Sheet Music[1] and the *Roman de la Rose* Digital Surrogates of Medieval Manuscripts.[2] Recognition and analysis applications are being built for these collections using our locally developed open-source *Gamera* framework. [3] In both cases, the usage patterns suggest that the needs of document image analysis technology are different from those would be served by transcription alone.

2. The Lester S. Levy Collection of Sheet Music

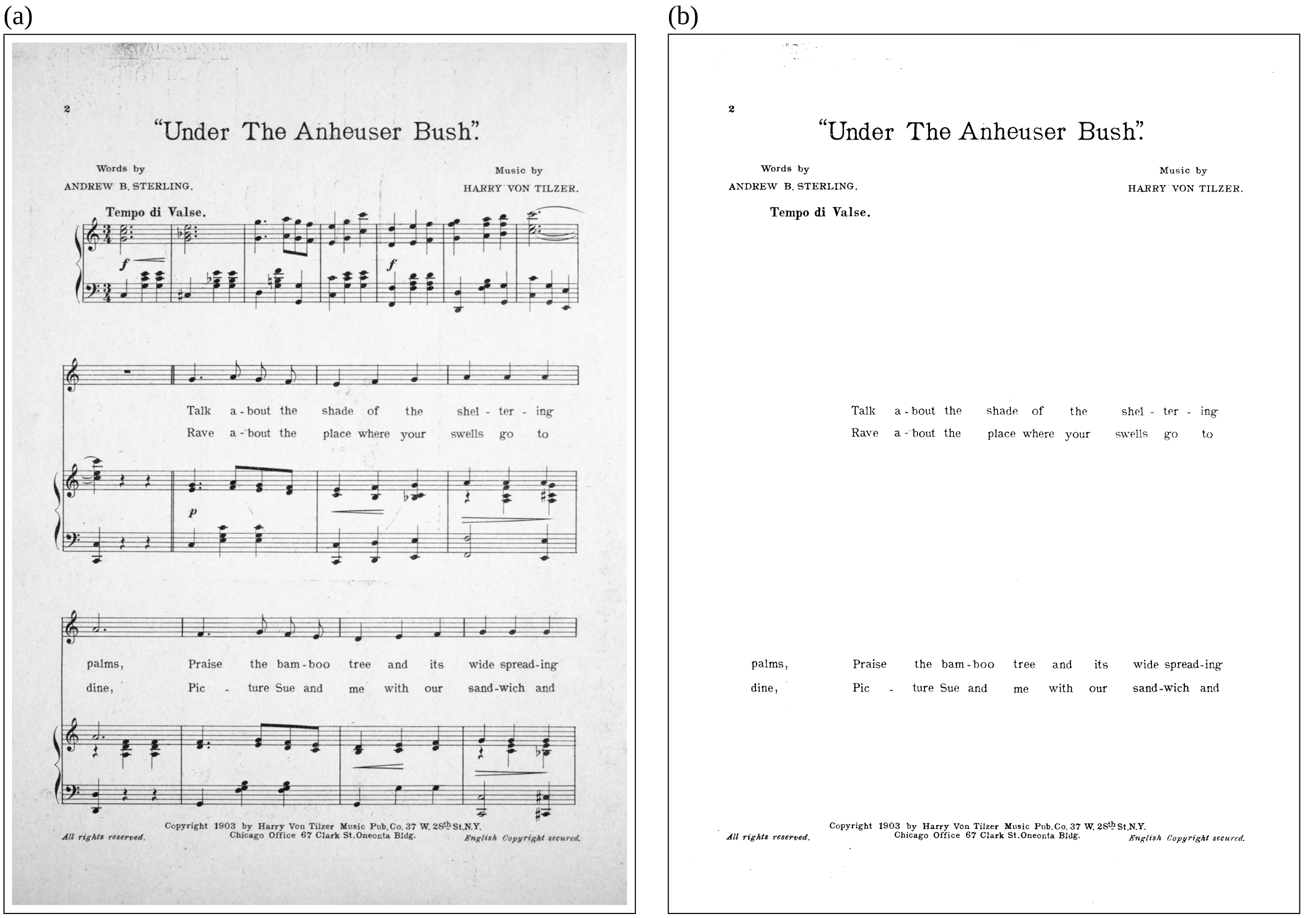
The Lester S. Levy Collection of Sheet Music, part of the Special Collections of the Milton S. Eisenhower Library at The Johns Hopkins University, comprises nearly 30,000 pieces of common Western music notation and associated cover art dating from the 18th to mid-20th centuries. The primary focus of the collection is American popular music and its strength lies in its illustration of American history through its commercial musical output.

Work on an Optical Music Recognition (OMR) system for the Levy Collection began at the Digital Knowledge Center (DKC) in 2000 [2], building upon work that began in the mid-1980's [4]. Developing a transcription system for sheet music covers many different areas of image processing and machine learning and combines the researchers' interests in computer science and music. Unfortunately, while this research is seen as a panacea by some and could potentially open the collection to new users, it is not necessarily relevant to the *existing* users of the Levy Collection. According to the collection's manager, most queries are not musical in nature. Instead, users of the Levy Collection tend to be more interested in the historical content, which is most concrete in the lyrics and cover art. With this realization, we felt it was necessary to pursue different applications of document image analysis to meet those needs in tandem with our own research interests.

2.1. Lyric extraction

Since the historical content of the music is most readily obtained from the lyrics, it would be useful to create a text-based representation of them. Such a format would permit

[1] `http://levysheetmusic.mse.jhu.edu/`
[2] `http://rose.mse.jhu.edu/`


(a)
(b)
2
"Under The Anheuser Bush".
Words by
ANDREW B. STERLING.
Music by
HARRY VON TILZER.
Tempo di Valse.
Talk a-bout the shade of the shel-ter-ing
Rave a-bout the place where your swells go to
palms, Praise the bam-boo tree and its wide spread-ing
dine, Pic - ture Sue and me with our sand-wich and
Copyright 1903 by Harry Von Tilzer Music Pub.Co.37 W. 28th St.N.Y.
Chicago Office 67 Clark St.Oneonta Bldg.
All rights reserved.
English Copyright secured.

(c)

Talk a bout the shade of the shel ter ing
Rave a bout the place where your swells go to

palms, Praise the bam boo tree and its wide spread ing
dine, Pic ture Sue and me with our sand wich and

(d)

Talk about the shade of the sheltering palms,
Praise the bamboo tree and its wide spreading charms,

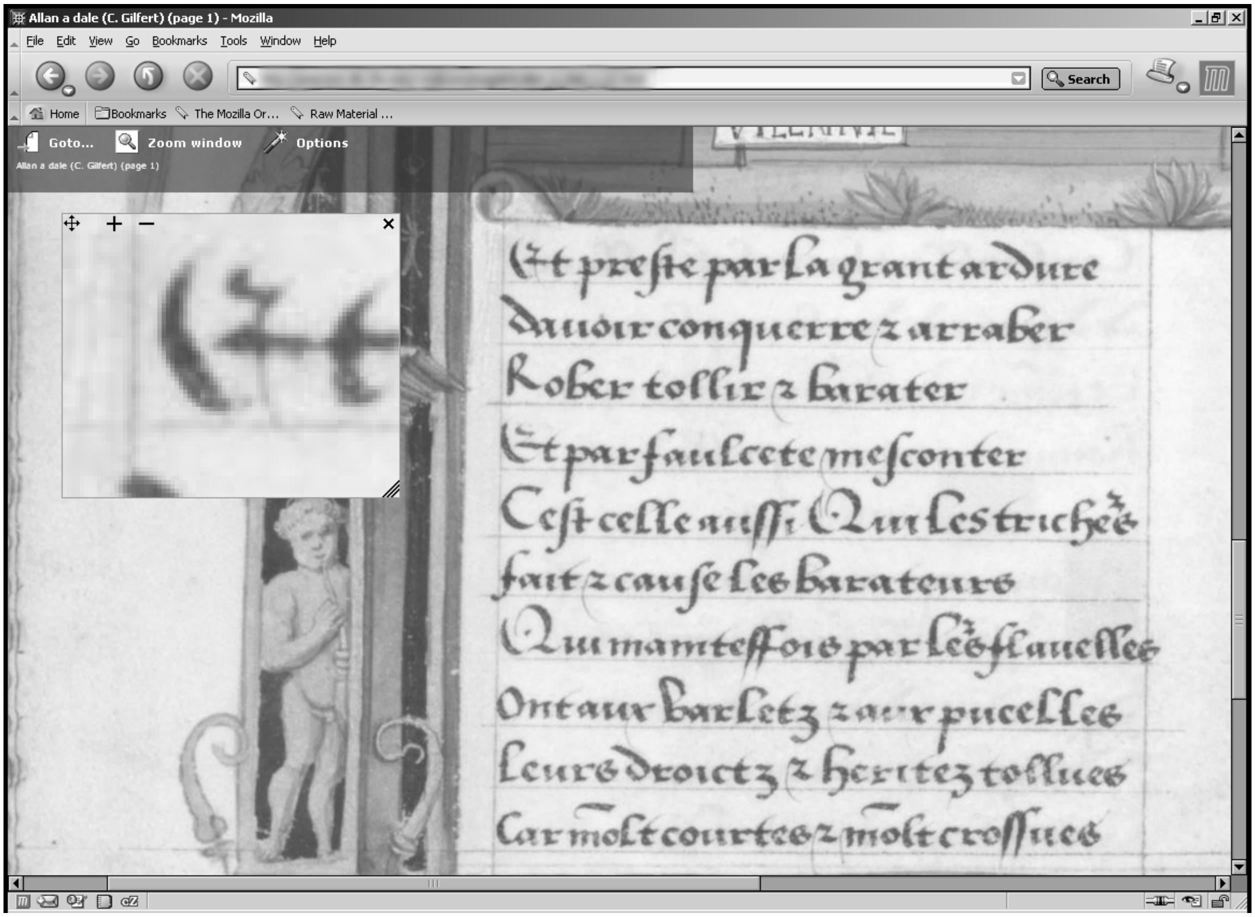
Rave about the place where your swells go to dine,
Picture Sue and me with our sandwich and stein,

Figure 1. "Under The Anheuser Bush" by A. B. Sterling and H. Von Tilzer (1903). (a) The original page; (b) The page with all the musical elements removed; (c) The lyrics as they appear after OCR; (d) The correct performance order and syllabic grouping.

full-text searching and other natural language processing-based analysis.

To extract the lyrics, the existing OMR system is first used to filter out all information that appears to be "musical" (Figure 1). Then, font- and language-independent OCR is applied to the text. Since words on a musical score are typeset with separated syllables, any system that relies on a dictionary for error correction does not perform well. Grouping the lyric syllables back into words is non-trivial, as the cues for the grouping of syllables (i.e. hyphens or underscores) are frequently inconsistent or absent, and many groupings can be ambiguous [9] [8, ch. 17]. There are also issues in restoring the lyrics to their sung order, particularly in strophic music.[3] (Figure 1, for example, has two verses of lyrics written to the same line of music). This cannot be solved without an understanding of the musical "flow control" symbols (e.g. repetition signs, *dal segno*, etc.) on the

[3] *Strophic:* A section of music where the same melody is performed multiple times using different lyrics.



Allan a dale (C. Gilfert) (page 1) - Mozilla
File Edit View Go Bookmarks Tools Window Help
Search
Home
Bookmarks
The Mozilla Or...
Raw Material ...
Goto...
Zoom window
Options
Allan a dale (C. Gilfert) (page 1)
+
−
×

Et preſte par la grant ardure
Dauoir conquerre z arraber
Rober tollir z barater
Et par faulcete meſconter
Ceſt celle auſſi Qui les trichee
fait z cauſe les barateurs
Qui maintefſois par leſ flauelles
Ont aux varletz z aux pucelles
Leurs droictz z heritez tollues
Car molt courtes z molt croſſues

Figure 2. The image concordance browsing environment. Clicking on a line of text pops up a menu that allows the user to jump to the corresponding line of text in another edition of the document.

page [8, ch. 12]. Fortunately, our pre-existing OMR technology is able to capture that information quite well.

3. The *Roman de la Rose* Digital Surrogates of Medieval Manuscripts

A joint project of Johns Hopkins University and the Pierpont Morgan Library,[4] this digital collection contains six manuscripts of *Roman de la Rose*, one of the most widely read works in the French language. If the project is successful, the collection may expand to include more of the over 300 extant copies available throughout the world.

One of the interesting features of the collection is that it is a parallel corpus: while the manuscripts are not identical, they are derivative works from a common source. This creates some interesting opportunities for information analysis and extraction that we hope to explore in the near future.

3.1. Image concordance browsing and annotation

For the medievalists we've talked to, the ability to obtain high-quality facsimiles of any page in the *Rose* collection in an instant is already a quantum leap ahead of the traditional level of access to the materials. However, a logical extension of this is to allow hyperlinks between particular regions on the page, which would allow the researcher to more conveniently navigate amongst related regions.

An image-based concordance browsing and annotation environment (Figure 2) is being developed that runs in most modern web browsers without the use of any plugins. When the user clicks on a "hotspot" in the image, a menu of other related hotspots is presented. Hotspots can also point to comments or other sources of information on the world

[4] With the participation of the Walters Art Museum, Bodleian Library and The J. Paul Getty Museum

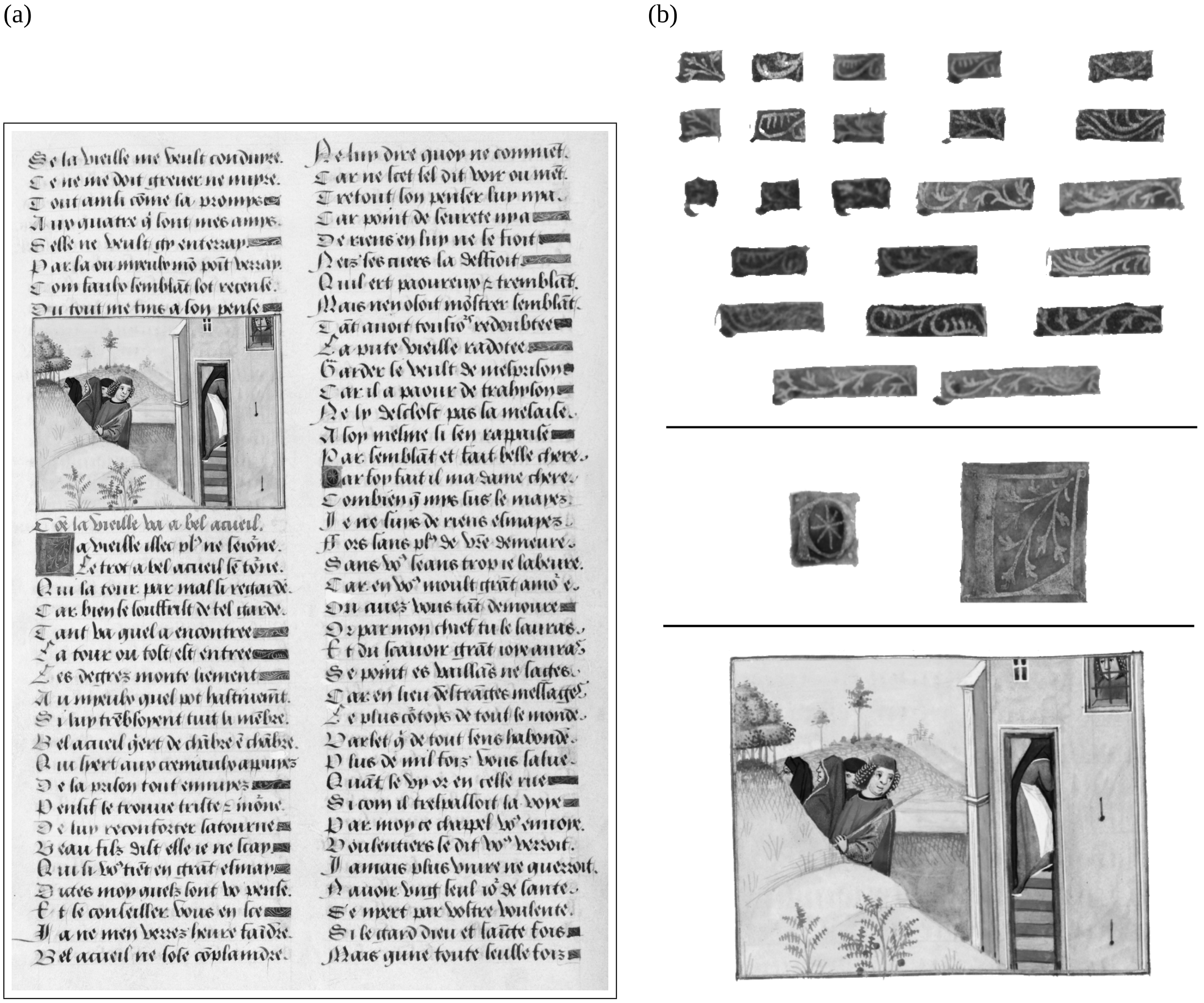


Figure 3. Folio 20 from the Douce manuscript in the *Roman de la Rose* collection showing automatic illumination extraction. (b) shows all of the illuminations extracted from (a), automatically grouped into three classes: decorations, characters and illustrations.

wide web. These hotspots and links could be statically pregenerated by a document image analysis backend, producing an *m*-to-*n* mapping of lines. However, as there will always be contention over such results (as there would be with any single human researcher as well), new hotspots and links can be added or edited by end users at any time. The author of each edit is recorded, so that different communities of authorities can develop and the tool can be used for scholarly discourse. In that way, the environment can be thought of as a fine-grained image-based annotation collaboratory [1, 5].

3.2. Illumination extraction

Of considerable interest to medievalists are the illuminations[5] in a manuscript, both their content and where they occur. While their ornamentation, and thus intra-class variation, makes them very difficult to recognize as specific characters automatically, their identification as ornaments as distinct from body text is quite simple, using only features of size and color variation (Figure 3). This automatic building of illumination "swatches" is something that has been requested by the medievalists and we will soon be in a position to evaluate their use for real research.

[5] *Illuminations*: Colorful illustrative or decorative elements.

4. Encouraging participatory design

Despite the preliminary success of the subprojects presented here, they still could benefit from a more formal approach to participatory design. In cooperation with our in-house usability specialist, we're beginning to explore ways of making this research more user-centric [7, 6]. One of the biggest obstacles faced thus far in the various collaborations has been mutual understanding of the relative difficulties of different kinds of research. It appears to be difficult for non-software developers to estimate the difficulty of certain development tasks. For instance, it is widely assumed, in the community at large, that OCR is a solved problem for many different kinds of documents. Conversely, it has been difficult for software developers to estimate the amount of time spent on various humanities research tasks. Determining which problems will have the most "bang for the buck" is a balancing act between their technical tractability (how long until it becomes usable?) and the amount of impact (how much time will this save, and how much will it enable?) For this reason, we are developing a framework in which to undertake more formal studies of the tasks in which document researchers spend their time. From these studies, we hope to determine which of those tasks are "menial", and thus good candidates for automation, and which are inherently "human". This information will help us further focus our research around the needs of collection users.

While this approach is clearly related to the library's mandate as a service to the University, it is also important to note that research in document image analysis can also bring collections to new users who previously would not have been able to meaningfully access them. This sort of long-range research should not be abandoned.

5. Conclusion

Examining user needs not only increases the relevancy of the technology to the collection users, but can also be a source of new and interesting research problems. Additionally, these problems are often more tractable in the short term than general transcription. We expect this approach to be a win for all involved.

References

[1] H. Brocks, A. Stein, U. Thiel, I. Frommholz, and A. Dirsch-Weigand. How to incorporate collaborative discourse in cultural digital libraries. In *ECAI2002 Workshop on Semantic Authoring, Annotation & Knowledge Markup*, 2002.

[2] M. Droettboom, I. Fujinaga, and K. MacMillan. Optical music interpretation. In *Structural, Syntactic and Statistical Pattern Recognition*, pages 378–386, 2002.

[3] M. Droettboom, K. MacMillan, and I. Fujinaga. The Gamera framework for building custom recognition systems. In *Symposium on Document Image Understanding Technology*, pages 275–286, 2003.

[4] I. Fujinaga. *Adaptive optical music recognition*. PhD thesis, McGill University, 1996.

[5] J. Kahan, M.-R. Koivunen, E. Prud'Hommeaux, and R. R. Swick. Annotea: An open rdf infrastructure for shared web annotations. In *WWW10 International Conference*, 2001.

[6] F. Kensing, J. Simonsen, and K. Bodker. MUST: A method for participatory design. *Human-Computer Interaction*, 13(2):167–198, 1998.

[7] M. J. Muller. Layered participatory analysis: New developments in the CARD technique designing with and for others. In *ACM Computer Human Interaction CHI Conference on Human Factors in Computing Systems*, pages 90–97, 2001.

[8] G. Read. *Music Notation: A Manual of Modern Practice*. Taplinger, 1979.

[9] B. Wingenroth, M. Patton, and T. DiLauro. En han cing ac cess to the Le vy Sheet Mu sic Col lec tion: Re con struc ting full - text lyr ics from syl lab les. In *Joint Conference on Digital Libraries 2002*, pages 308–309, 2002.

Acknowledgments

Thanks to Cynthia Requardt and John Spitzer for their feedback on usage of the Levy collection.

Thanks to Steve Nichols, Nadia Altschul and Rip Cohen for their feedback and helpful suggestions for the Rose collection.

And thanks to Teal Anderson, our resident usability specialist, for introducing me to the field of participatory design.

This research is funded in part by NSF grant #IIS0205466.