# Digitizing musical scores: Challenges and opportunities for libraries

Audrey Laplante
École de bibliothéconomie et des sciences
de l'information, Université de Montréal
CIRMMT
+1 514-343-6048
audrey.laplante@umontreal.ca

Ichiro Fujinaga
Schulich School of Music
McGill University
CIRMMT
+1 514 398-4535
ich@music.mcgill.ca

## ABSTRACT

Musical scores and manuscripts are essential resources for music theory research. Although many libraries are such documents from their collections, these online resources are dispersed and the functionalities for exploiting their content remain limited. In this paper, we present a qualitative study based on interviews with librarians on the challenges libraries of all types face when they wish to digitize musical scores. In the light of a literature review on the role libraries can play in supporting digital humanities research, we conclude by briefly discussing the opportunities new technologies for optical music recognition and computer-aided music analysis could create for libraries.

## CCS Concepts

• **Human-centered computing~User studies** • **Applied computing~Sound and music computing.**

## Keywords

Music digital libraries; digitization; musical scores; libraries; digital humanities; user studies.

## 1. INTRODUCTION

Humanities researchers have used computational approaches almost as soon as computers became available. The first studies on computer-assisted stylistic analysis and authorship attribution appeared in the early 1960s (e.g., [12]). Music researchers followed shortly after. At the end of the 1960s, Arthur Mendel and Lewis Lockwood from Princeton University had encoded several hundred pages of the works of Josquin Desprez on punched cards, allowing them to perform computer-based stylistic analysis and investigate composer attribution [11]. Technological advances, most notably the advent of the personal computer and, later, the World Wide Web, contributed greatly to the consolidation and the development of digital humanities [6]. As libraries of all types started to digitize documents in their collections, researchers gained access to new resources in digital format. Like books and articles, an increasing number of musical scores and manuscripts—the primary object of study of music theorists—are now available online. Unfortunately, some barriers still preclude music researchers from taking full advantage of these online collections. One problem is the

accuracy rate of optical music recognition (OMR) technologies, which is much lower than with optical character recognition (OCR) technologies. As a result, musical scores digitized by libraries are usually available as images only, without the possibility of searching their content. Collections of encoded music scores, such as those used for the Josquin Research Project[1] and MuseData[2], have most often been transcribed manually. Considering how time-consuming this is, only a small portion of musical scores is currently available in an encoded format [13]. This means that music researchers who want to use computational approaches for music analysis need to either choose from existing collections, or manually encode the music scores they wish to study. One other problem is the encoding format. While the *Text Encoding Initiative (TEI) Guidelines* were published in 1994, the Music Encoding Initiative framework was developed later [5]. The lack of a unified encoding framework for music has had a negative impact on the development of applications for music analysis.

The Single Interface for Music Score Searching and Analysis (SIMSSA) Project[3] aims to help lower these barriers by (1) developing OMR technologies, by (2) developing infrastructure for processing music documents to create a large collection of encoded music scores, and by (3) developing an interface with tools to search and analyze these documents. However, for this to work, libraries will have to continue to digitize their collections of musical scores, to create rich metadata, and to make the digitized images compliant with SIMSSA infrastructure.

To better anticipate what could hinder the adoption of the tools that are being developed by the SIMSSA team, we interviewed librarians with the aim of understanding what challenges libraries of all types face when they undertake or wish to undertake musical score or manuscript digitization projects. In the first part of this paper, we present the results of this study. In the second part, based on the literature on the role libraries could play in supporting digital humanities research, we briefly discuss the opportunities for libraries to get involved in supporting digital musicology.

## 2. CHALLENGES OF DIGITIZATION
## 2.1 Related work

Surveys and case studies on digitization reveal that libraries face many challenges when undertaking digitization projects. Funding is the most commonly mentioned challenge [1, 8-10, 14]. Digitization is expensive: it requires a lot of resources,

---

[1] http://josquin.stanford.edu/

[2] http://www.musedata.org/

[3] https://simssa.ca

especially in terms of staff time, as these projects usually involve preparing, handling, and scanning the documents, performing quality control, manually creating metadata, and providing access to the collections through a website and/or a search interface. One issue is that many digitization projects are financed with special funds, such as grants or gifts from donors, rather than with operating funds, which often means that a temporary team is assigned to the project rather than a well-established team [10, 14]. Related to that, another central issue that emerges is the lack of staff [9], particularly competent staff that can offer technical assistance [8, 9]. The long-term preservation of digital objects is another important issue [1, 10], which could also be tied to the fact that many projects are funded with special, short-term funds. As Breeding explains, ensuring long-term access to digital content entails "a constant and unbroken chain of migrations to new storage media, file formats, and controls to ensure and correct the integrity of all the bits that comprise each digital object" [1]. Other challenges include dealing with legal issues (mainly related to copyrights) [1, 10] and workflow management [9].

Similar issues arise from an examination of the case studies on the digitization of musical scores. Workflow management, however, appears to be particularly problematic [2, 3, 16]. Reasons that could explain this include the fact that musical scores are often multimodal (in addition to the symbolic notation, scores might include lyrics and images) [3]; some are composed of various parts, which complicates handling [16]; and the preprocessing steps can be tedious if technologies for the extraction of the symbolic content of the scores are used [2].

## 2.2 Methods

The objective of this study was to understand the challenges libraries confront when they carry out, or plan to carry out, a musical score digitization project. Since this is an exploratory study that aims to provide a *rich description* of these challenges from the librarians' perspective, a qualitative approach was considered most appropriate as a first step.

### 2.2.1 Data Collection

The data was collected through in-depth, semi-structured interviews with librarians working in various types of libraries. Two interview guides were developed: one for libraries that had already completed such a digitization project, and one for those who were still at the planning stage. The review of the literature on digitization in libraries provided a useful background for the development of the interview guides. During the interviews, participants were asked to describe their library's musical score digitization project, whether planned, ongoing, or completed. They were invited to talk about the issues that arose at different stages of the project. They were finally asked what they considered the greatest challenge was for libraries wishing to digitize musical scores.

### 2.2.2 Participants

An invitation to participate in the study was sent to IAML-L, the mailing list of the International Association of Music Libraries, Archives and Documentation Centres, in the spring of 2015, just before the annual conference of the Association that was held in New York, June 21-26. Seven interviews were conducted during the Conference or on the following days. Three more librarians who had shown interest in the study were contacted and interviewed in the spring of 2016, for a total of ten participants. Five participants worked in an academic library, three in a national library, one in a public library, and one in a conservatory library. They were from Belgium, Canada, the United Kingdom, and the United States. The interviews were conducted in English or French, according to the participant's preference.

### 2.2.3 Data Analysis

The interviews were audio-recorded and transcribed manually. Each interview lasted between 37 and 62 minutes. The data was analyzed using the qualitative data analysis software QDA Miner from Provalis Research. The transcripts were subdivided into units of meaning (i.e., into paragraphs). The data was analyzed inductively, using the conventional content analysis approach described in [7], an approach that is deemed suitable for descriptive studies. This method is iterative. The first step consists in reading through the data to derive an initial coding scheme, using the words of the participants. Following readings are used to develop thematic categories, identify the common characteristics of the units assigned to each category, and finally explore relationships between categories.

## 2.3 Findings

Collectively, the participants represented a wide variety of experiences in terms of musical score digitization. Two participants talked about projects that were still at the planning stage; two were managing large-scale digitization projects, and the others were managing small- to medium-sized projects. The projects also covered a variety of document types and periods: music manuscripts, sheet music, scores for large ensembles, for chamber music, and for solo instruments, from the medieval period to the 20th century, from popular and folk music to classical music. Two participants were female, but to maintain confidentiality, the masculine form is used in this paper to refer to the participants, regardless of their gender. The following sections present the most common challenges mentioned by the participants.

### 2.3.1 Short-Term Funding and Planning Issues

Participants complained about lack of funding. When they managed to find some money to get their digitization project off the ground, it almost always came from a special fund. One participant talked about an endowment fund that "they were supposed to spend every year", suggesting that an unexpected surplus had been allocated at the last minute to his project. One talked about a two-year "special grant from the university" and another about a "summer grant". One also deplored that, although they used to have money for this sort of project, they now had to "look for outside funding to get anything done."

Relying on special, short-term funds for these projects is not without consequences. Perhaps the most striking finding of this study is that in libraries where digitization projects were ongoing or completed, most participants reported having started to digitize the documents before having carefully planned the entire project. Two participants explained that they did not know how they would provide access to the digital collections when they started the digitization. One said: "I haven't quite worked out all the details for getting it supported for a public interface; I've got to get that going," adding later, "I guess that's a problem that I don't really know what platform it's going to be on." Similarly, another said: "I haven't quite figured out how I want to represent the materials online." Access was not the only thing they had not planned for. One participant's answer suggests that the whole workflow had not been entirely developed and tested at the outset of the project: "That means we have to put all these pipes with multiple F numbers in certain cells […]. I think I'm not sure this is really gonna work." He then added that he saw this as "just the start" and was hoping he could plan for something more ambitious "down the road". Another explained that the lack of planning, more specifically

the fact that they had neglected to involve the cataloging department in the planning process, had created a conflict within the organization (which he referred to as a "political situation"), and a bottleneck that had caused the digitization process to be suspended. He concluded by saying that it had been "a learning experience".

### 2.3.2 Creating Metadata: A Difficult and Time-Consuming Task

Metadata was another central issue that arose from the interviews with the librarians. While two participants explained that traditional MARC cataloging records were used, others felt that they needed to provide not only high-quality but richer-than-normal metadata, and had thus devoted great attention to this aspect. A participant explained that he preferred digitizing less documents to make sure he had the resources needed to provide enriched records: "We decided […] we'd rather have a small number of items and have a lot of metadata." He then added that they were including "a lot of contextual metadata" to the records. Similarly, a participant said that, according to him, "you need really, really good metadata," because "when you put things out, you need to do it responsibly, in a good way, and be able to use it properly". Another participant explained that they were enriching MARC records with "any kind of numbers that have been assigned to [the] works," as well as "movements", "performing forces", and "kind of subject words" that consisted mainly in music genres and forms. Not surprisingly, creating these very detailed records represents a challenge. One considered that "dealing with the accompanying cataloging and metadata requirements" was the greatest challenge he had faced with regards to the musical score digitization project he was managing, as it was "more of a burden than the scanning itself". As will be seen in the next section, this was closely linked to staffing issues.

### 2.3.3 Lack of (Skilled) Staff

Questions regarding staffing elicited similar comments from participants: all complained about not having enough staff to conduct digitization projects the way or at the pace they desired. It was considered by all but one participant as the greatest challenge associated with conducting musical score digitization projects. One talked about the "lack of human resources", and another about the difficulty of getting the required "people power". But most of the time, it was the lack of *skilled* staff the participants deplored, particularly for creating metadata and for providing technical support. A participant indeed explained that there was no problem with the scanning process—"Everyone knows how to do it"—, the problem was with creating "good metadata". Although the participants recognized the need for assigning skilled staff to this task, the fact that the digitizing projects were mostly financed with special funds meant that they often had to hire temporary, untrained staff, rather than trained catalogers. In academic libraries, students were almost invariably in charge of creating metadata. A participant reported that it was done by "a large team of student employees", and another explained that a "humanities computing student" had received "a summer grant for it".

Technology support was the other area where the participants felt they did not have access to the required expertise. A participant indicated that he "need[ed] the technical support". For another participant, the biggest challenge was to find staff that had the technical expertise while being able to understand the musicological aspects: "If I express what I need, they should be able to translate it into the best of the technological solutions." For him, this was mainly a communication issue: "You have people that have the musicological knowledge and you have people who have the technical knowledge; they don't necessarily connect with one another or they might not have the common vocabulary." But to digitize and give access to musical scores, librarians have to work not only with IT staff but also with other services and, of course, with music researchers.

### 2.3.4 Limited Collaboration with Researchers

Most participants indicated that they were maintaining regular contact with music researchers, from inside and outside their institution, who could be interested in the score collection they were digitizing. These activities, however, consisted mainly in promoting the collection through presentations in conferences of music societies or associations, or through articles in music library or musicology journals. One explained that he was regularly going to conferences of relevant music societies to "make a presentation about the availability of [the collection]". Another was planning "to start trying to write papers and give conferences" about their new online score collection. A participant had a slightly different strategy. In addition to writing an article in a music library journal, he indicated that he was considering "trying to get graduate students or faculty to work with the collection […] so that it gets promoted at conferences or in published articles."

However, apart from these promotional activities, the participants did not report having worked closely with music researchers. Researchers had not been consulted, at least not formally, during the planning process, for instance to make decisions regarding the metadata or the user interface. That does not mean the planning had been done without the end-user in mind (the efforts they reported investing in metadata shows they cared about end-users), but it was mostly based on their informal knowledge of the needs of music researchers. For instance, one participant indicated that the search interface had been "designed specifically to help music researchers", although there had been no consultation or usability testing.

### 2.3.5 The Challenge of Using OMR Technologies

All musical score digitization projects described by participants gave access to digital images only. This means that the content of scores was not searchable. When asked why they had not used OMR technologies to extract the symbolic notation, most participants expressed doubts regarding the accuracy of such systems. One said: "I've never heard about a really good solution for character recognition for music." Related to that, another participant was anticipating he would have to correct the errors, which would be costly: "that's something that I suppose is not as efficient yet in terms of cost/effectiveness [as OCR]." Another reason that arose from the participants' responses was that most were outsourcing the scanning of the scores to third-party service providers, and although the providers typically performed OCR on textual documents, OMR was not offered. A participant also expressed that he did not feel he knew enough about OMR technologies to use them: "I don't know enough about that, how well OMR can work and, you know, can our vendor or whoever is doing digitizing do that?"

### 2.3.6 Summary of Findings and Limits

The interviews with the librarians revealed that libraries face various challenges during the planning and the implementation of musical score digitization projects. Many of these challenges were not specific to musical scores: as seen in Section 2.1, the lack of long-term funding and skilled staff, as well as the challenges associated with metadata creation, were also identified as important issues by librarians managing other types of digitization projects. Digitizing musical scores, however, also seem to bring its own challenges, especially at the technical

level: OMR technologies are less reliable and are not as easy to use as OCR, which explains why none of the librarians interviewed reported using OMR to make the musical scores searchable. It was also noted that music researchers did not seem to have played an important role in the planning process of these projects. Participants seemed to consider these projects mostly as promotional tools for the library and/or the institution rather than as a way to strengthen the collaboration between the library and the music researchers. However, copyrights did not seem to represent a major challenge in most cases as the musical scores that were being digitized had been, for the most part, in the public domain for a long time.

Considering the small size of the sample, the results presented here should be interpreted carefully. Dealing with a small sample allowed us to obtain rich descriptions from each participant, which is useful at an exploratory stage. The next step will be to conduct a survey that will be widely distributed through the IAML listserv in order to verify the generalizability of our findings.

## 3. OPPORTUNITIES FOR LIBRARIES AND RESEARCHERS

Most musical scores that are available online are not searchable, which limits greatly the possibilities for researchers who want to use computational approaches to music analysis. As already mentioned, OMR technologies are not as advanced and easy to use as OCR yet, which means that to extract the music notation of digitized images, libraries would need to allocate time and resources. Should libraries play a more important role in providing music researchers with the resources and tools they need to adopt these computer-based approaches? Many librarians think so and have thus called for a closer collaboration between digital humanities researchers and libraries (e.g., [4, 15, 17]). As mentioned in the introduction, because of the limited availability of encoded music scores, music researchers often have to create their own corpus of encoded scores as well as their own tools. This poses several challenges that libraries could help surmount or alleviate. Libraries could help create these corpora: they own large musical score collections and have expertise in digitization. They would, however, need to start using OMR technologies. Libraries also have the infrastructure and the expertise to help ensure the long-term preservation of these corpora. Indeed, if researchers can sometimes get funding for creating these corpora, it is much more difficult to secure funding for their maintenance. Entrusting the management of a digital corpus to the library would therefore increase the probability that it remains readable beyond the end of a research project. Other areas of libraries' expertise that could be called for within the context of a digital musicology research project include metadata creation, user interfaces, and dissemination/outreach.

Libraries would also benefit from such collaborations. A library that would provide online access to encoded music scores as well as tools for searching and analyzing these scores would increase its visibility, both within its community and outside. The role and place of the library have been evolving during the last two decades due to technological advances. In an era where an increasing proportion of information resources is available online, people are sometimes questioning the relevance and value of libraries. Taking a bigger part in digital scholarship represents an opportunity for libraries to redefine their role and reaffirm their relevance. Hence, Vandegrift and Varner suggest that libraries should "increasingly function as a place where scholars can try new things, explore new methodologies and generally experiment with new ways of doing scholarship" [17]. Although this implies a change in the type of services usually offered by libraries, there is an obvious compatibility between these services and the values of the library profession. The tools that are being developed by the SIMSSA teams (e.g., the OMR technologies and the interface for searching and analyzing musical scores) should help libraries play a more important role in digital musicology scholarship.

## 4. ACKNOWLEDGMENTS

## 5. REFERENCES

[1] Breeding, M. Ongoing challenges in digitization. *Computers in Libraries*, 34, 9 (Nov. 2014), 16-18.

[2] Choudhury, G. S., Requardt, C., Fujinaga, I., DiLauro, T., Brown, E. W., Warner, J. W. and Harrington, B. Digital workflow management: The Lester S. Levy Digitized Collection of Sheet Music. *First Monday*, 5, 6 (Jun. 2000).

[3] Damm, D., Fremerey, C., Thomas, V., Clausen, M., Kurth, F. and Müller, M. A digital library framework for heterogeneous music collections: from document acquisition to cross-modal interaction. *Int J Digit Libr*, 12, 2 (Aug. 2012), 53-71.

[4] Fay, E. and Nyhan, J. Webbs on the Web: libraries, digital humanities and collaboration. *Libr. Rev.*, 64, 1/2 (2015), 118-134.

[5] Hankinson, A., Roland, P. and Fujinaga, I. The Music Encoding Initiative as a document-encoding framework. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)* (Miami, FL, USA, Oct. 24-28, 2011). University of Miami, Miami, 2011.

[6] Hockey, S. The history of humanities computing. In *A companion to digital humanities*, S. Schreibman, R. G. Siemens and J. Unsworth, Eds. Blackwell Pub., Malden, MA, 2004, 3-19.

[7] Hsieh, H.-F. and Shannon, S. E. Three approaches to qualitative content analysis. *Qual. Health Res*, 15, 9 (Nov. 2005), 1277-1288.

[8] Institute of Museum and Library Services. *Status of technology and digitization in the nation's museums and libraries, 2006*. Institute of Museum and Library Services, 2006.

[9] Lampert, C. and Vaughan, J. Success factors and strategic planning: Rebuilding an academic library digitization program. *Inform. Technol. Libr.*, 28, 3 (Sept. 2009), 116-136.

[10] Lopatin, L. Library digitization projects, issues and guidelines: A survey of the literature. *Libr. Hi Tech*, 24, 2 2006), 273-289.

[11] Mendel, A. Some preliminary attempts at computer-assisted style analysis in music. *Computers and the Humanities*, 4, 1 (Sept. 1969), 41-52.

[12] Mosteller, F. and Wallace, D. L. Inference in an authorship problem. *J. Am. Stat. Assoc.*, 58, 302 (June 1963), 275-309.

[13] Pugin, L. The challenge of data in digital musicology. *Frontiers in Digital Humanities*, 2(Aug. 2015).

[14] Rieger, O. Y. Enduring access to special collections: Challenges and opportunities for large-scale digitization initiatives. *RBM: A Journal of Rare Books, Manuscripts and Cultural Heritage*, 11, 1 (Mar. 2010), 11-22.

[15] Sula, C. A. Digital humanities and libraries: A conceptual model. *J. of Libr. Admin.*, 53, 1 (Jan. 2013), 10-26.

[16] Tobar, C. Music to my ears: The New York Philharmonic digital archives. *D-Lib Magazine*, 17, 7-8 (Jul. 2011).

[17] Vandegrift, M. and Varner, S. Evolving in common: Creating mutually supportive relationships between libraries and the digital humanities. *J. of Libr. Admin.*, 53, 1 (Jan. 2013), 67-78.