Digital Humanities at Oxford Summer School (DHOxSS) 2015: Digital Musicology Workshop

by Sean Luyk

The [Digital Humanities at Oxford Summer School (DHOxSS)](#) is one of a handful of professional development opportunities[1] offering training to students and researchers on the issues, methods, and scholarship in the growing field of digital humanities (DH). DHOxSS has an innovative format that combines thematic, hands-on workshops with a common program of keynote lectures, optional events, and opportunities for networking. In July 2015, thanks to generous support from my institution, I was able to spend a week at Oxford University to attend the Digital Musicology workshop stream of DHOxSS. I was joined by a mix of students and early-career faculty from institutions across the world, all either already engaged in digital humanities related music projects, or, like myself, interested in developing their skill set further. My experience in the Digital Musicology workshop stream of DHOxSS has proven to be quite beneficial, and has resulted in my active participation in a number of DH-related music projects, and in a developing knowledge base that has allowed me to speak more confidently on these topics. I was left feeling optimistic about the leadership role that music librarians can play in supporting and developing digital research in music, and excited about what the future may hold. Given the increasing interest in the digital humanities by music librarians,[2] and the scholarly and professional benefits I have drawn from attending DHOxSS, I thought I would share some of my experiences in this piece for the *CAML Review*, organized into broad themes by method or concern.

Music Information Retrieval (MIR): Methods, Tools, Projects

The DHOxSS brought in a number of well-known experts in the field of MIR, and therefore a significant portion of the workshop dealt with what are widely considered to be MIR research techniques and projects. Born from information retrieval (IR) research of the 1960s, and

Sean Luyk (sean.luyk@ualberta.ca) is the Music Librarian at Rutherford Library, University of Alberta.

1. Such as the Digital Humanities Summer Institute held annually at the University of Victoria: http://www.dhsi.org/
2. For example, a THATCamp was held at the 2015 MLA annual conference: http://musiclib2015thatcamp.org/.

developed by researchers in a wide range of fields from musicology to computer science, current work in MIR primarily deals with music as auditory events. As a result, much of the MIR portions of the workshop focused on methods and tools for analyzing music as audio, and on defining what musicological insights can be drawn from these methods. J. Stephen Downie provided participants with a solid theoretical background on the implications of MIR for musicological research, and outlined the significance that the [International Society for Music Information Retrieval (ISMIR)](#) has had on the development of MIR as the dominant paradigm in computational methods applied to musicology. Downie also outlined the research methods which are of primary importance and use in MIR, namely: audio feature extraction, machine learning, and classification. Participants were given ample time to gain experience working with all of these research methods, and the format was a combination of lectures, followed by hands-on sessions guided by experts at a comfortable pace.

Christophe Rhodes and Chris Cannam led us in audio feature extraction techniques using [Sonic Visualiser](#), [AudioDB](#), and a number of useful [Vamp](#) (audio feature extraction) plugins, to make determinations of musical similarity based on various musical features. This was followed by additional hands-on practice using programmatic methods to analyze the data gathered from feature extraction using the [Python programming language](#), accessed through the [IPython notebook](#) (now known as the [Jupyter Notebook](#)). As the number of tools listed suggests, entering the field of MIR research is a daunting task. However, there is a large and active community of MIR researchers out there who are willing to help, and all the tools used are open access, meaning that documentation and a community of users are not hard to find.

J. Stephen Downie, Ben Fields, and Tillman Weyde introduced participants to the fascinating area of machine learning as it is applies to music information retrieval research. Machine learning techniques are of primary importance to the study of music as auditory events, with significant research in this area coming out of Queen Mary University of London, the Music Technology Group at the Universitat Pompeu Fabra in Barcelona, McGill University, and others. Many of the gains made in MIR research over the last 25 years are a result of the adoption of machine learning research methods, and it was clear from this portion of DHOxSS that anyone interested in engaging in MIR research needs to become familiar with machine learning at some point. MIR research that uses machine learning techniques follows a standardized protocol for the testing of algorithms and the verification of MIR systems,[3] which means that the results of research are objectively and independently verified. Downie cautioned workshop participants that the research results gleaned from machine learning techniques are probabilistic, and not deterministic. What this means in practice is that a single, "correct" outcome regarding a

---

3. http://www.music-ir.org/mirex/wiki/MIREX_HOME

musical feature (e.g. tempo estimation) is not to be expected; instead, what is accomplished is a level of certainty along a sliding scale.

Machine learning falls under two broad approaches: supervised and unsupervised. In supervised machine learning, it is necessary to supply a set of human-created test (or "ground-truth") data which the machine learning algorithms use for comparison. For example, the SALAMI project, which made great progress in automated formal analysis, used music theory graduate students at McGill University as the source of ground-truth data, which was in turn used to train the machine learning algorithms in predicting musical form. Unsupervised machine learning does not use any test data, and the machine learning algorithms attempt to label/cluster data using specified features. We were given hands-on practice using the unsupervised approach. We began with a data set of sound recordings, and conducted audio feature extraction using jAudio. Audio features were then analyzed using the open source machine learning tool Weka. In addition to practice using this multi-tool approach with relatively small data sets, we were also introduced to the web-based Digital Music Lab, a powerful tool for doing MIR work on some very large and significant sample data sets, including the CHARM and I Like Music collections.

Learning more about the MIR research process has helped me to further develop my capacity to think computationally about musicological and music librarianship questions. I've been prompted to start asking different questions of the researchers I work with, of the materials I curate, and in my own research program. How can one think of a musicological question in a data driven way? How can computers help to automate the tasks of music librarians? How can we work to improve our search and discovery tools so that they better meet the needs of music information seekers and users? These are just some of the questions that I've been pondering since attending the MIR portions of DHOxSS.

Music Encoding: MEI, MusicXML, and Music21

If MIR is primarily about studying music as audio, approaches from computational musicology are about studying music as represented in symbolic form. In the symbolic domain, much higher order musical features are encoded than in the audio domain, which means that the insights from this approach are perhaps more focused towards questions of music theory and analysis. This section of the DHOxSS program dealt with the tools and research concerns that relate to studying notated music using computational methods. Tim Crawford presented his study of the lute repertoire, and the resulting database, Electronic Corpus of Lute Music (ECOLM). Crawford outlined the process of using optical music recognition (OMR) software to digitize lute scores using Gamera and Aruspix, and the ongoing process of encoding this repertoire to make it machine readable. Participants were given hands-on experience using the

two dominant music encoding schemes: [Music Encoding Initiative (MEI)](#) and [Music XML](#), as well as a Python toolkit for doing music analysis with encoded music, [music21](#). The potential that a world of fully encoded music could bring to music research was demonstrated in Ichiro Fujinaga's presentation on the [Single Interface for Music Score Searching and Analysis (SIMSSA) project](#). The SIMSSA project aims to create a "Google Books for scores," to provide worldwide access to scores from a single interface. Much like optical character recognition (OCR) in Google Books has transformed digitized books into texts searchable online, the SIMSSA project will use OMR technologies to do the same for music scores. The key pieces of this project are: [Rodan](#) (an OMR workflow engine), Gamera (staff removal and correction), Aruspix (OMR software for early music), [Verovio](#) (a music engraver), and [Neon.js](#) (for neume generation).

Data

Much attention was given at DHOxSS to thinking about music as data, both as audio (as summarized in the section on MIR), but also as data *about* music, which includes metadata, linked data, and performance data. There were unfortunately no hands-on sessions devoted to this topic, but participants enjoyed a number of lectures outlining some interesting research projects that attempt to use big data as a way to study music. Studying music from a big data perspective in many ways flips the current dominant paradigm of doing micro-level, highly detailed histories of music, which in sum, present comprehensive histories. This is not to say that big data approaches will replace detailed histories. Instead, thinking about big data prompts scholars to ask new questions, as it reveals relationships and insights that might be missed by more closely focused methods.

Stephen Rose outlined the [Big Data History of Music project](#),[4] which took a "distant reading"[5] approach to the study of published music in the holdings of the British Library, manuscripts inventoried in RISM, and concert life as evidenced in the [Concert Programmes database](#). Rose and his team looked at these massive data sets (in excess of 5 million records) to see if traditional music histories centred on canonical composers could be challenged, and if new narratives could be pieced together. The researchers were attempting to do "macrohistory," and were successful in revealing peripheral histories of music through their efforts. For example, their approach allowed them to identify publishing centres outside of the typical list of cities, and clusters of repertoire not previously considered to be significant in the history of music. This project used bibliographic data created over many years, so data cleanup was a

---

4. See also: Sandra Tuppen, Stephen Rose, and Loukia Drosopoulou, "Library Catalogue Records as a Research Resource: Introducing 'A Big Data History of Music'," *Fontes Artis Musicae* 63, no. 2 (2016): 67-88, accessed April 18, 2016, https://muse.jhu.edu/.

5. Distant reading is a term and research approach developed by Franco Moretti in his book *Distant Reading (*London: Verso, 2013).

significant aspect of this work. Rose and his team exported MARC records from the British Library catalogue using MarcEdit, and did substantial cleanup using OpenRefine. Similarly, Rachel Cowgill's presentation on InConcert discussed the challenges and rewards of studying concert life from a data perspective. The InConcert project aims to take a data-based approach to the study of concert life as evidenced in printed ephemera sources such as programmes, bills, and reviews. Some new scholarly principles are emerging given the realities of this type of scholarship, which is very exciting.

Kevin Page introduced us to the potential that linked open data (LOD) has in representing musicological knowledge on the web. Page's session was primarily an in-depth tutorial on the basics of the Resource Description Framework (RDF), ontology creation, and the data that needs to be created to achieve true linked open data. Page also outlined the work that needs to be accomplished in order for musicology data on the internet to move from a web of documents to a web of data. Carolin Rindfleisch presented some promising preliminary work that demonstrates what can be done with annotation data from live performance. Rindfleisch, a scholar of Wagner, and in particular of the leitmotif, presented a toolkit for annotating live opera performance. Based on her interest in creating a data set and ontology for studying leitmotifs in Wagner's work, Rindfleisch first provided us with an overview of her study of leitmotifs, existing scholarly work on the subject, and the challenges of studying this topic. Rindfleisch then described a pilot project in which she annotated a live performance of a Wagner opera using a tablet interface and an Echo smartpen, making live annotations that could later be automatically imported and matched with the recording of the opera.

Broad Concerns

One of the main advantages of the way DHOxSS is organized is that the individual, thematic workshops are complemented by keynotes and guest lectures that typically deal with broader concerns of interest to all workshop participants. Akin to the *Big Thinking* lecture series that takes place at the annual Congress of the Humanities and Social Sciences, the keynotes and individual lectures at DHOxSS are designed to bring together all participants to experience presentations related to research that asks critical questions and reports on current issues. The following will summarize some of the highlights from last summer's sessions of this nature.

The first keynote of DHOxSS 2015 was by Dr. Jane Winters, Professor of Digital History and Head of Publications in the Institute of Historical Research, School of Advanced Study, University of London. Winters' talk, "How I Learned to Stop Worrying and Love the Digital"[6] focused on the challenges and opportunities faced by humanists (especially historians) in

---

6. Winters' talk can be viewed here: https://podcasts.ox.ac.uk/how-i-learned-stop-worrying-and-love-digital.

conducting research using web archives and other forms of big data as primary sources. According to Winters, we're at a point where the interest in and importance of using these data in research exceeds the knowledge and tools available to adequately do so. As Winters explained, humanists are not necessarily used to approaching their research from the macro level, and distant reading approaches are considered suspect by some. However, rich data sets documenting recent histories are increasingly becoming available, and it is important for scholars to be able to gain insight from them.

One project that exemplifies this issue is the Big UK Domain Data for the Arts and Humanities project, of which Winters is a primary investigator. The project attempts to both develop a theoretical and methodological framework for studying the massive, 65 terabyte UK domain web crawl that took place between 1996 and 2013, and develop tools to find meaning in it. Progress is being made on this front, but there is much work to be done before web archives of this size can be effectively used for research. Winters cited some earlier projects, including the Digging Into Linked Parliamentary Data project, and the large collaborative project, the Early English Books Online Text Creation Project (EEBO-TCP), as examples of longstanding digital humanities projects that have met the formidable challenge of dealing with large amounts of (previously) unstructured information and making it useful to researchers and readers alike. Winters closed her talk with a call for action that will ring true for music librarians and musicologists alike—the need to develop better tools and techniques for dealing with time-based media—an issue that will take some time to address adequately.

The Digital Transformations Panel brought together four seasoned DH researchers from various disciplines,[7] to discuss how research, teaching, and learning can be better situated to respond to the potential that the changing digital landscape presents.[8] Andrew Prescott noted that the digital transformation, rather than being a linear path, is more a series of subtle changes with no continuum of long-term development. Digital change needs more time for exploration and play; we need to be more open to sharing and remixing in our research and teaching, and be willing to experiment, and even to fail. Tim Crawford focused his talk on the importance of collaboration in the digital humanities. Crawford, a musicologist by trade, has benefited over the years by collaborating with researchers with more technical backgrounds, rightly stating that it isn't realistic to try to learn all the skills necessary to get a successful digital project off the ground. Jane Winters echoed Tim Crawford's point about collaboration, calling for more collaboration between researchers and librarians, a direction I'm sure many *CAML Review*

---

7. David De Roure, Oxford e-Research Centre, University of Oxford (Chair); Lucie Burgess, Bodleian Libraries, University of Oxford; Tim Crawford, Computing Department, Goldsmiths, University of London; Andrew Prescott, University of Glasgow; and Jane Winters, Institute of Historical Research, University of London
8. The Digital Transformations Panel can be viewed here: https://podcasts.ox.ac.uk/digital-transformations.

readers would support. Winters also focused much of her talk on the issue of infrastructure, especially as it relates to dealing with born-digital materials. Infrastructure problems are only going to get worse, and if the digital transformation in research is to happen, the infrastructure deficit will need to be addressed. Winters also spent part of her talk on the topic of ethics and anonymity in digital research, commenting on "unexpected findings" and the need to protect individuals identified as a result of research using web archival sources. Lucie Burgess' talk dealt with research data management, and the importance of developing better tools for humanists in this regard.

The closing keynote by James Loxley, entitled "Uneasy Dreams: The Becoming of Digital Scholarship"[9] was one of the highlights of DHOxSS 2015, as it helped me to frame what I was learning and experiencing all week from a broad perspective.[10] Loxley, Professor of Early Modern Literature at the University of Edinburgh, gave a thoughtful talk discussing the rise of digital humanities research, and illustrated how we can find longstanding humanistic values and methods inherent in the digital turn. It is essential to understand the differences among being digital, doing digital, and becoming digital; emphasis should be placed on the last, and care must be taken in disciplinary responses to the rise of digital scholarship. In this regard, Loxley cautioned against repeating history, and used the rise of critical theory in the 1960s and 70s as an example of this tendency. Situating digital scholarship as the new critical theory, Loxley warns us not to go "post-al"; there is no "post-digital" phase that we are to expect at any moment—rather, we are *becoming* digital. Loxley used Ernest L. Boyer's influential four dimensions of scholarship as a way to frame digital scholarship,[11] and argued that each of the four dimensions have yet to catch up to the changes wrought by the digital revolution. Loxley left the audience feeling optimistic about the future; digital humanities is about so much more than just creating tools, it is about changing what scholarship means, while staying true to humanistic values.

---

9. Loxley's talk can be viewed here:http://podcasts.ox.ac.uk/uneasy-dreams-becoming-digital-scholarship.
10. Loxley created a BuzzFeed listicle for audience members to follow, which can be found here: http://www.buzzfeed.com/jamesloxley31/my-talk-for-dhoxss-1pkln.
11. Boyer's four dimensions of scholarship include discovery, integration, application, and teaching. See Boyer, E. L. (1990), *Scholarship Reconsidered: Priorities of the Professoriat,* Carnegie Foundation for the Advancement of Teaching, https://depts.washington.edu/gs630/Spring/Boyer.pdf.