Eleanor Selfridge-Field

Search engines for digitally encoded scores

As dehumanized as computer-assisted research is often perceived to be, no account of a subject such as this one can be entirely value-neutral. Despite a legacy of more than half a century of effort, the rapid evolution of technology has forced the development of digital tools for score search and analysis, which is of little interest to industry, to fend for itself through countless changes of computer operation and endless rounds of data migration. These comments are based mainly on involvement in our work (from 1984) at the Center for Computer Assisted Research in the Humanities (CCARH); our graduate courses in computer-assisted music topics at Stanford University (from 1994); various collaborative involvements with multiple spokes of the RISM project (principally from about 2000); and our in-house projects MuseData (from 1984) and Themefinder (from 1996). The last two named now have progeny of their own, which are mentioned below. A list of web links can be found at the end of this article.

For musicians the term 'music search' embraces diverse phenomena, for digital technology magnifies the differences between written music, recorded music and our burgeoning wealth of scanned materials. In the commercial arena 'music search' means audio search. It is heavily supported by commercial interests seeking market advantage.[1] To librarians it often means access to scanned resources. This discussion addresses scores that have been fully encoded by one of the many methods developed for the purpose. Most users will have some familiarity with music notation software, which utilizes encoded data but may not make it easily separable from the score itself. These visible products have many parallels in academic and non-profit precincts.

Why might we want to be able to search musical content? Finding resources is a central reason, and while non-specialist users may be satisfied with a title-match to any version of a widely known work, specialist users often want a specific edition. Textual metadata may be adequate, if most of the particulars are known. They work less well if the user knows the music well but bibliographical details are elusive. Score searching is far from perfected, and many of the lapses lie less in the software used than in the nature of music itself. For music historians and analysts, music search promises far more than the mere finding of a printed score that corresponds to a piece recalled. Searches of entire repertories for a particular *cantus firmus*, a pre-existing melody, a basso ostinato or an intentional paraphrase of an earlier piece are all legitimate possibilities. Studies of source filiation have been aided from time to time by close comparison of musical content in diverse copies of the same work. Studies of authorship likewise look to details of musical fabric for what psychologists call 'unconscious details' (particular intervallic patterns or note sequences, comparative lengths of melodic lines, the tessitura of individual parts) of a score. These kinds of features pass unnoticed in normal music listening. The musical examples discussed here are familiar ones intended to entice new users to explore on their own, not to answer specific questions.

Dynamics of music searches

The digital world often depends on ranked lists to steer users to what they want. Ranking can be based on a single dimension, but music is inherently multidimensional. Ranking is too simplistic to accommodate varying weights for pitch, duration, lyrics, articulation, timbre and so forth. If two pieces have the same metre but are in different keys or modes, they could be the same or different. It would depend on several other factors. Yet two 'pieces' that have the same title and a common composer are rarely identical in every respect.

Early Music, Vol. XLII, No. 4 © The Author 2014. Published by Oxford University Press. All rights reserved.
doi:10.1093/em/cau099, available online at www.em.oxfordjournals.org
Advance Access publication October 9, 2014

Related to this superficial approach to culling pieces that are possibly related to one another is the need for nuance in the construction of a music search. Most search software in current use is one-dimensional: pitch is privileged over all else. Within that one dimension it may be possible to seek with greater or lesser detail. Looser queries (ignoring, let us suppose, key or mode and chromatic inflection) normally produce a higher yield but one with more invalid matches among them. Yet a search that is too literal may miss a valid match that is mistranscribed or slightly varied for reasons of adaptation. Finding the optimal approach in any given situation can usually benefit when it is possible to 'fine tune' individual parameters of the music. For example, allowing pitch patterns to be general while specifying a key (or vice versa) may improve the likelihood of finding what is sought. Joint searches of pitches and rhythm have been shown to be highly effective but are rarely supported because they are difficult to implement.

Pitch is the most common feature by which encoded music is searched because it is the most apparent. Written music pitch has three parameters—a note name (A, B, C, D, E, F, G), an inflection (for example F♯, B♭) and an octave number. The level of detail desired by the user must be supported both by the underlying encoding and by the search software. Musicians can easily be deceived about the capability of the tool in use because they are conversant with a written or aurally remembered example of the music sought. Pitch-search capabilities that are limited to letter names constitute a diatonic or 'base-7' scheme. The roster of names for a full octave allowing for single sharps and flats on any tone requires 21 differentiated positions (7 diatonic pitches × 3 chromatic states); a roster accommodating double sharps and flats requires 35 such positions (7 diatonic pitches × 5 chromatic states).[2] Officially the octave ascending from *c′* is the fourth, but in electronic systems octave numbers are unpredictable.[3]

The ubiquity of MIDI-enabled keyboard instruments has imposed another layer of pitch nomenclature on the tones of the octave through the assignment of 'key numbers', which serve to address physical keys rather than pitches per se. MIDI's complete subservience to equal temperament creates certain problems in addressing the needs of pre-tonal music, since the lack of a sounding difference between C! and D*y* produces an error-prone path to enharmonic transcription. It is a long-standard practice in MIDI-based software to describe the black note between A and B as 'A!'.[4] To define pitch correctly in chromatic contexts (through single sharps and flats) would require a base-21 representation, and for tonal repertories from the early 18th century onwards a base of 35 (or more) pitch positions per octave is safest.

Search engines for music

The search engines discussed here are related to two large clusters of projects—those originating and/or maintained at the Center for Computer Assisted Research in the Humanities (CCARH) at Stanford University, and those related to the Répertoire International des Sources Musicales (RISM). These constellations have entirely different purposes but both support search technology for encoded musical data. MuseData, focused on the development of means to encode, print and archive full scores to a scholarly standard, operates on in-house software developed by Walter Hewlett. Since 1984 a team of data specialists has encoded roughly 1,200 works from 1680 to 1850, including significant quantities of orchestral and chamber repertory plus selected operas and oratorios. All data can be downloaded as is and a few hundred pdfs of scores (some with accompanying parts) are also downloadable. Most of the data has been translated repeatedly into diverse formats for the purposes of re-editing, analysis or online presentation.

Themefinder, which stores musical incipits of classical and folk repertories totalling slightly more than 100,000 items, was initiated by David Huron, a CCARH visitor in 1996 and subsequent years. His aim was to monitor how users go about searching for music. Since that time the search capability has been expanded as the repertory has grown. The sample repertories it includes come from classical works from 1650 to 1900, Latin motets, and folksongs from Europe and Asia. Five levels of search are supported: (1) pitch (A...G plus differentiated sharps and flats); (2) intervals (including direction); (3) scale degree (1...7); (4) gross contour (up, down, repeat); and (5) refined contour (up/down by step or leap, or repeat).[5] Themefinder also has filters for metre and mode.

The Josquin Research Project (JRP), under the direction of Jesse Rodin, with technical implementation by Craig Sapp, is a project of recent vintage (2010) that currently holds more than 550 works, including roughly 100 multi-movement settings of the Mass Ordinary. Its search engine is adapted from Themefinder to better suit scores in mensural notation. It is possible to search by pitch, interval and rhythm, to scan tessituras, to find relative usage by tone, and to see visual summarizations of rhythmic usage over the course of an entire movement. Filters for genre and mensuration are among those currently supported. Holdings for Josquin are separated according to whether or not a work's attribution is secure. The second largest number of works is by Pierre de la Rue, followed by pieces by Ockeghem, Dufay and others.[6] Since the project is ongoing these numbers are destined to increase.

The RISM project, the origins of which can be traced to the 1950s, coordinates music bibliography of many kinds. The most visible project is the music manuscript inventory ('A II' in the original nomenclature), which has been searchable online since 2011. Its aim is to inventory all manuscripts containing music from the 17th and 18th centuries throughout the world. It originated (as did MuseData) long before the internet, but RISM was conceived from the start to be a computer-based project. This was in the mainframe era, when digital collaboration involved exchanging physical records. Much of the musical incipit data was first transcribed manually and later encoded. The manuscript inventory is rich in text fields, but our focus here is on its one field for encoded musical incipits.[7] It was an early hope of the founders that the music-transcription field would facilitate the ready identification of anonymous works.[8] Other projects under the RISM umbrella include 'A I' (printed music entirely by one composer) and 'B I' (printed anthologies of music), which have somewhat different profiles of development.

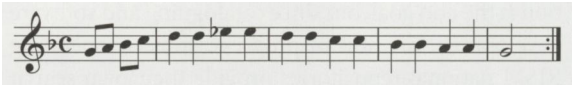
The RISM manuscript inventory is based on the transcription code Plaine & Easie, which was developed in 1966 by Barry Brook and Murray Gould.[9] Although the format has been modified and the data translated many times, the encoded incipits remain reasonably transparent. At this writing, the central collection serves almost 900,000 listings. The completion of the composite (international) collection managed by the RISM editorial office in Frankfurt, with online access facilitated by the Bavarian State Library in Munich, is still some years away. Full synchronization is the end goal, but since cataloguing and software development are still in progress, some cooperating RISM national repositories provide their own search engines. These include those of Ireland, Switzerland, the United Kingdom and the United States (hereafter RISM EI, RISM CH, RISM UK and RISM US).

The small differences between them are quite instructive to those interested in further exploring the myriad possibilities and pitfalls of music search. In 2004 CCARH set up, with the cooperation of the central editorial offices and the RISM US office, an experimental website for searching US music manuscripts. Based on the Themefinder search engine, it preserves Themefinder's multi-tiered search functions but also supports text search of some principal fields. This example inspired a virtual-keyboard adjunct to Laurent Pugin's RISM CH search site, parts of which were later incorporated in the RISM UK website. Pugin's music search page replaces Themefinder's search boxes with sliders for pitch and duration.[10] It also has filters for metre. Since many medieval manuscripts have been included in the Swiss RISM website, this ordinary fielded search supports searches by liturgical feast and other fields appropriate for such repertories.

Sample searches

A test set of incipits was assembled for the purpose of exploring the capabilities of these various search engines. Direct comparisons are not possible. The fact that there are few overlaps in data holdings means that some kinds of queries are destined to work in only one of the situations. The six themes chosen are grouped into three categories: (1) Renaissance dances, (2) authored polyphony (Ockeghem, Josquin) from the Renaissance and (3) late Baroque music (Handel, Bach). Some Renaissance dances had a long afterlife that carried through the 18th century. In a study that was nothing short of heroic, Luigi Ferdinando Tagliavini assembled a large collection of instances of the 'Ballo di Mantova' (ex.1).[11] Simple as the melodic outline is and quick though hearers are to find in it an antecedent of Smetana's *Má Vlast* (1879) and the Israeli national anthem (adopted in 1948), Tagliavini's manual cull of loose matches, which now ranges far beyond 100 examples, is concentrated in the 16th to 18th centuries. It is not difficult to find among them

Ex.1 An early example of the 'Ballo di Mantova' (RISM OPAC)



pieces that deviate by key, metre, mode or rhythmic detail in ways that some would consider to invalidate the match. Yet the study is enormously valuable for demonstrating that a determined human being may outperform a computer search by a very large distance. In most of them a musical ear will hear a melody similar to Smetana's piece.

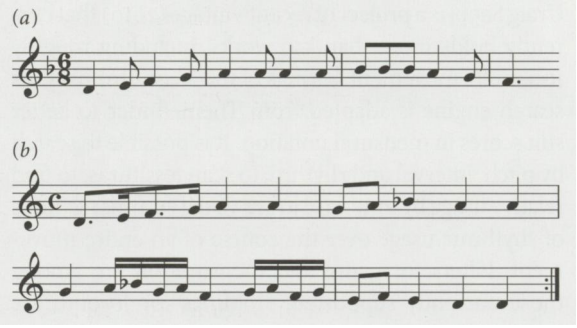
A note-by-note search for the first nine pitches of the 'Ballo di Mantova' in the RISM OPAC (international) database located only eleven instances and, as with most melodic searches, produced some hits of questionable value. A melodic search is especially useful in cases such as this where titles are almost as numerous as instances of the melody. In a scale-degree search with a minor-mode filter and two wildcards in a nine-note profile Themefinder found 30 matches, some of which fail for rhythmic, metrical or implied harmonic reasons, as shown in ex.2a–b.

Musicians familiar with other tune families of the 16th and 17th centuries will appreciate that the 'Ballo di Mantova' phenomenon (and its attendant search problems) has numerous parallels in such popular items as 'La Girometta', the chaconne bass, La Folìa and others of similarly broad dispersion.

Renaissance polyphony

Searching for Renaissance polyphony raises different issues. It is a widely held view and the foundation of generations of research that much of the period's more elaborate music, especially sacred vocal music, is based on earlier secular songs and includes melodies once associated with particular texts, often in tightly knit schemes of rhyme and repetition. The degree of matching can revolve in the polyphonic context on other kinds of detail, among them mensuration, number of voices and rhythmic properties (which can vary greatly from one part to another). Some of this kind of variance can be judged from a comparison of Ockeghem's three-voice song 'Ma Maistresse' (ex.3) and his four-voice Mass.

Ex.2a–b Two 'fuzzy' matches for the 'Ballo di Mantova' melody from Themefinder



When we turn to the start of the Mass setting we find a significant redeployment of voices such that they are treated in pairs and in a more clearly organized imitative manner. However the new discantus interleaves elements of the old discantus voice with the old contra voice (ex.4).

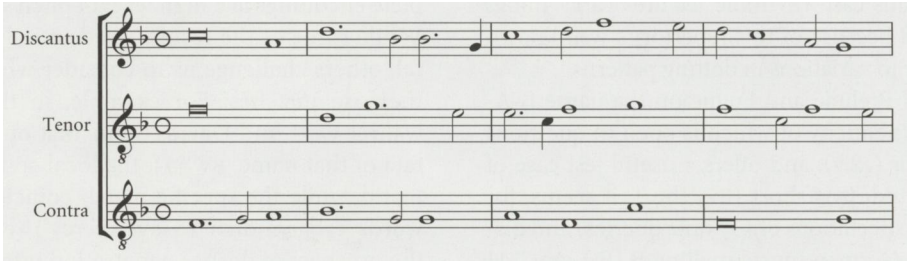
This comparison barely scratches the surface of an important implication for automatic melodic searches: not only may the line sought wander from part to part but it may also decay and be recomposed. Searches for melodic matches may confront a new battery of obstacles.

However, other music of the period presents much more straightforward possibilities. An almost contrary example is Josquin's well-known *soggetto cavato* 'Hercules Dux Ferrariae', which he used as the basis for a Mass; the eight tones of the subject were derived by vowel substitution (ex.5). A search through the JRP located eleven instances of its use within this Mass (they are heavily concentrated in the Sanctus). In searches elsewhere, RISM OPAC provides one quasi-match by Cristóbal de Morales in a Latin motet for the Annunciation. The melody seems not to have been one that natural powers of invention could have conjured up. The 'L'homme armé' melody, in contrast, elicits 175 matches in 71 works through its intervallic search (with the expression 1 4 1 -2 -2 -2).

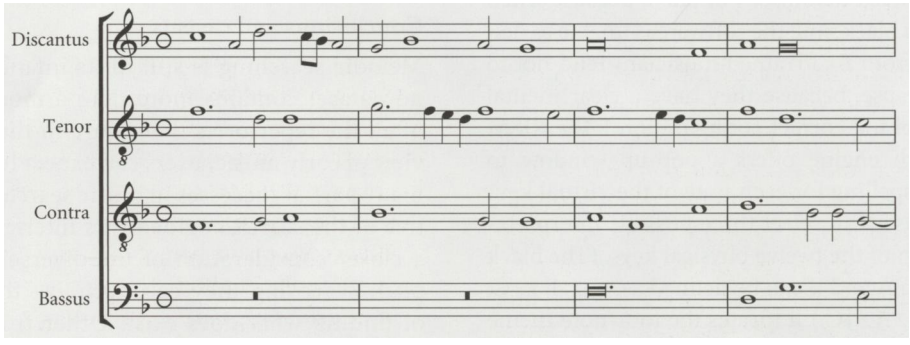
The Baroque

To represent the Baroque era we selected contrasted search examples—the aria 'Lascia ch'io piango' from Act 2 of Handel's *Rinaldo,* and the Bach fugal subject B-A-C-H. The aria is a

Ex.3 The beginning of Ockeghem's three-voice song 'Ma Maistresse' (Josquin Research Project)

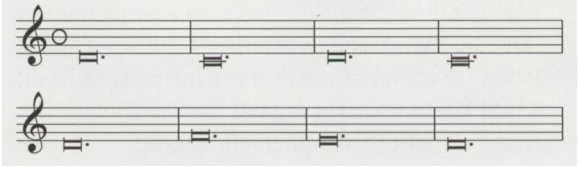


Ex.4 The beginning of Ockeghem's four-voice Mass based on 'Ma Maistresse'



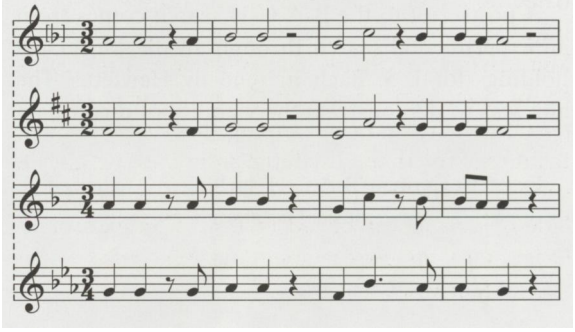
rhythmically simple diatonic one, while the Bach *soggetto cavato* is not simply chromatic but may require enharmonic definition in some search contexts. The difficulty one may encounter in searching for the aria is that because of its enduring popularity all the liberties of Baroque interpretation have accrued to its performing history over three centuries. Although historians generally see music as fixed to its time of origin and therefore to an authoritative version to produce legitimate matches, the 19th century produced far more musical manuscripts than earlier centuries and modifications were rampant. A text search on the aria in RISM OPAC cited among other sources two early ones,[12] one from later in the 18th century[13] and one from the 19th.[14] The subtle differences are easy to identify in comparison.

Ex.5 The *soggetto cavato* from Josquin's Mass 'Hercules Dux Ferrariae' (Josquin Research Project)



The surprising truth is that in a melodic search, 'Lascia ch'io piango' (ex.6) yielded the lowest hit rate across all the search engines mentioned here.[15] Because the singing of Baroque arias still varies as much today as it did in Handel's time, pitch searching that is too faithful to any one iteration is likely to miss others that are legitimate. Although the addition of rhythmic

Ex.6 Four versions of 'Lascia ch'io piango' from Act 2 of Handel's *Rinaldo*, as rendered in a series of 18th- and 19th-century manuscripts (RISM OPAC)

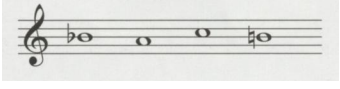


search features is a widely held goal of music search engines, in this case rhythmic features carry things further afield because of this repertory's well-known susceptibility to variations in dotting patterns.

The 'Bach' Prelude and Fugue on the name B-A-C-H (the authenticity of which is open to question) is in B♭ major (ex.7), and offers a useful test case of a different kind. It is short (too short, it seems, for RISM UK's expectations of 5–7 note queries) and distinctive. The German music spelling B [B♭]–A–C–H [B♮] proves to be as difficult to locate with virtual keyboards as without them. Virtual keyboards usually rely on MIDI definitions of pitch, because when a user plays a black note the software is unable to distinguish A♯ from B♭. (Trained musicians tend not to realize this lapse, because they have a clear mental conception of the correct configuration.) The RISM OPAC search engine offers a pop-up window to refine pitch spelling for each note of the virtual keyboard. The user can select three possible interpretations for each of the twelve physical keys. (The black note between A and B can be defined as 𝄪A / ♭B / 𝄫C; the B♮ as B / 𝄪A / ♭C.) It locates the four-note theme in 149 works, but the number of hits drops rapidly as the theme is extended (a phenomenon common to all melodic searches). Among its more likely matches is a theme in 3/4 from Act 3 of Kurt Stiegler's parody *Der Thomaskantor* (1928) on a text by F. A. Geissler. RISM US, utilizing the Themefinder search engine, finds five examples, among them the same J. L. Krebs 'Fuga alla breve' (Yale School of Music Library) as RISM reports in the Benedictine abbey of St Boniface (Munich) and the Prussian State Library (Berlin). In addition to RISM US, the Themefinder search engine[16] is also used in RISM CH, in the Josquin Research Project, and in a forthcoming project by Emiliano Ricciardi concerned with madrigal settings of verse by Torquato Tasso.

A postscript to the B-A-C-H search comes from an experimental search through all the MuseData holdings for J. S. Bach in 1996 by Hewlett.[17] The aim was to demonstrate the viability of direct data searches in encoded data, in this case given complete encodings at a high level of pitch specification. While some of the results were entirely coincidental, others challenge us to consider whether or not they are matches. For example, in the duet 'Du wahrer Gott und Davids Sohn' that opens the cantata of that name, BWV23, the local context is chromatic, while the specific match coincides with the words 'erbarm dich' ('Have mercy') (ex.8). Clearly the prospect of deeper penetration into polyphonic repertories will perplex us with further questions of interpretation.

Ex.7 The German 'B-A-C-H' spelling in 'Anglicized' musical notation



Conclusions and caveats

Melodic searching is still in its infancy and since no dataset contains more than a modest portion of all the repertories that figure in the wider purview of early music, users can expect to experience many partial successes in music searches, irrespective of the search engine or its interface. Through a closer consideration of the diverse methods of pitch searching they can optimize their chances of finding what does exist. Other frontiers loom on the horizon—methods of metrical and rhythmic search, coordinated pitch and lyrics searches, and support for seeking musical features pertinent to particular repertories. The frontiers in greatest need of user exploration and feedback lie in the particulars of Renaissance and Baroque methods of polyphonic imitation and melodic variation.

Music-search websites (open-access)

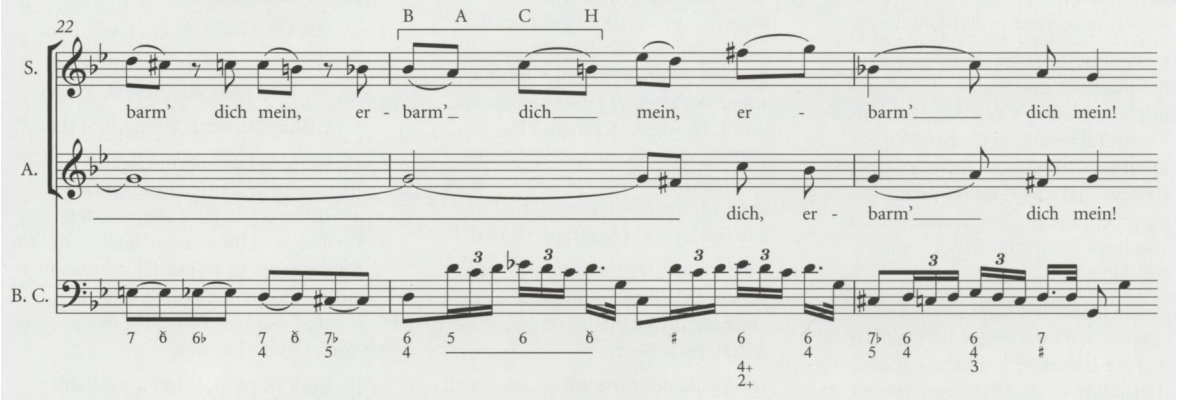
CCARH-maintained websites:

CCARH: www.ccarh.org
Josquin Research Project: http://jrp.stanford.edu
KernScores: http://kern.humdrum.org
MuseData: http://musedata.stanford.edu/
Themefinder: http://themefinder.stanford.edu/

RISM-related websites:

RISM CH: http://rism-ch.ch
RISM OPAC: https://opac.rism.info/metaopac/start.do?View=rism
RISM IE: www.rism-ie.org/ (no incipit search yet)
RISM UK: www.rism.org.uk/
RISM US: http://rism.themefinder.org

Ex.8 An incidental (?) match of the 'B-A-C-H' subject in Bach's cantata *Du wahrer Gott und Davids Sohn*, BWV23, at the words 'erbarm dich' (have mercy), from a MuseData full-text search



Teaching websites on musical search and analysis at Stanford University:

http://wiki.ccarh.org/wiki/Music_253
http://wiki.ccarh.org/wiki/Music_253/CS_275a_Syllabus
http://wiki.ccarh.org/wiki/Music_254
http://wiki.ccarh.org/wiki/Music_254/CS_275b_Syllabus
http://kern.humdrum.org

Eleanor Selfridge-Field, consulting professor of music at Stanford University and a researcher at the Center for Computer Assisted Research in the Humanities (an affiliate of the Packard Humanities Institute), is the author of six books and many articles in historical musicology as well as the editor of 16 yearbooks in digital musicology. She is a past member of the US RISM Committee and the international RISM advisory board. esfield@stanford.edu

Craig Sapp has kindly assisted in the preparation of the musical examples. He, Laurent Pugin, Jesse Rodin, Luigi Ferdinando Tagliavini, Ilias Chrissochoidis and the anonymous reviewers provided many useful comments.

1 Audio search can produce results that are more refined for the details of timing in recordings but far less competent in efforts to identify melodies, themes, harmonic patterns or procedures. Much of the audio-search software in current use relies at least partly on metadata (text fields such as title, author, performer and total length of the 'track').

2 Walter B. Hewlett's base-40 system goes one step further by inserting five null tokens (empty slots) to preserve correct enharmonic spellings in automatic transposition and harmonic assessment, as described in his 'A base-40 number-line representation of musical pitch', *Musikometrika*, iv (1992), pp.1–14, reproduced at www.ccarh.org/publications/reprints/base40/, and in his 'Method for Encoding Music Printing Information in a MIDI Message', U.S. Patent 5,675,100 (7 October 1997). The MIDI implementation later became known as MIDIPlus and is described in E. Selfridge-Field, *Beyond MIDI: the handbook of musical codes* (Cambridge, MA, 1997); reproduced with permission at http://beyondmidi.ccarh.org/beyondmidi-600dpi.pdf.

3 The US follows the convention of the American Acoustical Society (1939) in which the octave ascending from c′ is considered the fourth one. In MIDI documentation c′ = 60 (cited as c4), but some applications and manufacturers ignore this. Yamaha instruments refer to c′ as c3. Peachnote (http://peachnote.com), for example, currently produces playback that is an octave higher than the written pitch produced by its virtual keyboard, as c′ appears to be set to 72 (to generate the correct pitch, users can reset c′ to 60)

4 This would not affect searchability in a hexachordal repertory insofar as a B♭ could consistently be pursued as an 'A♯', but such usage does defy conventional terminology in discussions of both pre-tonal and tonal music. Over time commercial notation programs such as Finale and Sibelius (among many others) have developed algorithmic

solutions that remedy a high percentage of the ensuing errors inherent in the keyboard capture. They also provide menus to rename individual notes.

5 Most of the software in current use has been designed and implemented by Craig Sapp. Andreas Kornstädt designed the original interface. Roughly 20 former Stanford students have worked on various aspects of the project. Datasets have also been contributed by various colleagues.

6 The Dufay holdings have been contributed by Alejandro Planchart and are currently in process of data-translation to the JRP music format.

7 Other RISM projects focus on printed music by single composers, anthologies of printed music, music theory and so forth.

8 The last time its efficacy was assessed, the results had greater implications for dis-attribution than for attribution. See J. Schlichte, 'Der automatische Vergleich vom 83,243 Musikincipits aus der RISM Datenbank: Ergebnisse—Nutzen—Perspektiven', *Fontes artis musicae*, xxxvii/1 (1990), pp.35–46, and J. Howard, 'Strategies for sorting melodic incipits', *Melodic Comparison: Concepts, Procedures, and Applications, Computing in Musicology*, xi (1998), pp.119–128.

9 The most recent description is by J. Howard, 'Plaine and Easie Code: a code for music bibliography', in *Beyond MIDI*, pp.326–72.

10 Available at the advanced search page www.rism-ch.org/manuscripts/search?strategy=index after clicking 'incipits'.

11 L. F. Tagliavini, 'Il ballo di Mantova, ovvero, Fuggi, fuggi da questo cielo, ovvero, Cecilia, ovvero...', in *Max Lütolf zum 60. Geburtstag: Festschrift* (Basel, 1994). Tagliavini's collection is now being maintained and expanded by Liuwe Tamminga.

12 GB-Ob Mus. Sch. c. 41, GB-Lam Ms. 90.

13 US-NYp Mus. Res. jog 72–138.

14 I-BRc 19th cent, Fondo Pasini 1.

15 RISM OPAC identified 25 instances in a title search. With a wildcard search Themefinder turned up a loose relationship to two movements in a Handel harpsichord suite, and with a metre filter it located a copy of the aria in the New York Public Library.

16 Built from tools freely available at http://extras.humdrum.org/man/themax.

17 W. B. Hewlett, 'A derivative database format for high-speed searches', *Computing in Musicology*, x (1995–96), pp.131–42.