Eleanor Selfridge-Field (Stanford, CA)

The *MuseData* Electronic Corpora

When the Center for Computer Assisted Research in the Humanities (CCARH), a nonprofit research enterprise, was established by Walter B. Hewlett in 1984, one of its principal goals was to form electronic corpora of musical repertories from the 17th, 18th, and 19th centuries. Only people with a big appetite for unsolved problems could be attracted to such a goal. There was no MIDI (for musical keyboard entry of data) or internet (for distribution); e-mail was prohibitively expensive; and personal computers, although known, were not yet in widespread use. Even in the more modest field of text-processing, ASCII (the simplest code for representing letters and numbers in a computer) was not fully accepted. Systems which could print the standard diacriticals of European languages (à, ä, è, é, ì, Z, ⊥ et al.) were inordinately costly and unwieldy.

We took our inspiration from the example of classicists, who, over the preceding decade, had encoded most surviving ancient Greek literature. They did it by developing their own operating system, software, encoding schemes, and hardware. The IBYCUS computer[1] handled texts in Greek, Hebrew, and Coptic scripts as well as the Roman alphabet (including modern European diacriticals). While we did not use this computer at CCARH, we did use a variant of its operating system developed by Hewlett. It was particularly well suited to ›string‹ processing (i. e., the kind of data-searching people frequently do in literature, when they look for ›strings‹ of words of letter).

Goals and Methods

The goal of encoding musical scores into computer code required the development of many additional capabilities. The most primary need was that of facilitating the printing of music, because proofreading remains one of the most necessary and laborious elements in the processing of corpus formation. In order to print music, it is necessary first to encode it in some manner. There are many ways to do this, none of them perfect or complete, for there is no defined upper limit to the number of musical symbols. A related problem is that many musical symbols give approximate or ambiguous meanings which are clarified by visual context. For example, a dot lengthens or shortens duration, depending on its position relative to a note-head.

Two important aids to the early development of our own work were a survey of earlier work done in the fields of music printing and music representation over the preceding 20 years, and our convening in 1987 (at the invitation of the International Musicological Society) of the IMS Study Group on Musical Data and Computer Applications. This led

1 The IBYCUS desktop computer, developed by David Woodley Packard for applications involving ancient languages.

to a study of music-representation systems for the 20 years prior to the establishment of the Center and resulted, with the formation and participation of the IMS Study Group on Musical Data, in the anthology *Beyond MIDI*[2], in which 39 systems for representing music and manipulating musical data (for printing, sound, storage, and retrieval) are examined. Many milestones and new initiatives are also reported in the CCARH yearbook, *Computing in Musicology* (1985–). Titles in recent years have included *Melodic Similarity* (vol. 11, 1998), *The Virtual Score* (vol. 12, 2001), *Music Query* (vol. 13, 2004). Titles in production include *Music Analysis East and West* (vol. 14, 2006) and *Harmonic and Rhythmic Analysis by Computer* (vol. 15, 2007).[3] In all of these publications we have attempted to bring the best of computer technology to bear on the tasks designed and implemented by musicologists, with a view to assuring that long-term needs are adequately met and short-term goals are not overwhelmed by the latest ›trend‹ or gadget.

From the beginning, we have been sensitive to the problems of source ›philology‹ and the opportunities that computers afford for enabling the presentation of multiple versions of a musical work (a ›virtual score‹ in current parlance). Projects which have put down their roots in recent years have been able to carry these aims further than we have, for they build on capabilities made which in many cases have only become available within the past few years. New capabilities pose challenges as well as opportunities, however, because this year's fashionable procedure is often next year's dud. Compared to the number of decades or generations it has typically taken to produce one ›Gesamtausgabe‹, the lifetime of a technical capability can be very brief. We do something quite unfashionable, which we feel will not be superseded anytime soon: we encode all music in the simple alphanumeric code called ASCII. Each line of code (a ›record‹) provides space for annotations.

From this ›plain vanilla‹ master-set of data, which we call *MuseData*, many permutations can be derived by further processing. Conversion to the MIDI format produces files which can be used in pedagogical and rehearsal software, as well as for ordinary playback and proof-hearing of encoded materials. Conversion to a graphical image permits the assembly of scores, parts, and short scores. Conversion to the purely descriptive code used in the open-source, cost-free *Humdrum Toolkit* (lacking details of sound output or notational layout) enables the data to be examined for analytical purposes. (Nearly a hundred *Humdrum* tools currently exist.) To guarantee adaptability to the continuously evolving protocols for musical data, *MuseData* can be converted to and from *MusicXML*, an interchange code addressing individual components of music and musical layout. Via *MusicXML*, *MuseDat*a can be created from and read into popular notation programs, such as *Finale*® and *Sibelius*®. It is also directly translatable into SCORE, an industrial strength notation program used in the creation of numerous ›Gesamtausgaben‹ (Verdi, Wagner, Schoenberg, Ives, e. a.). Andreas Kornstaedt has been developing software to overlay personal annotations on a screen display of a score. Many other uses of the data are conceivable.

In our view of the universe, musical data which represents works of the distant past should be regarded as a springboard for virtual editions. If we look back on the mindset

2 Eleanor Selfridge-Field, *Beyond MIDI. Handbook of Musical Codes*, Cambridge and London 1997.

3 The series is published by The MIT Press (Cambridge and London).

of a century ago, when the preparation of many ›Gesamtausgaben‹ was vigorously pursued and the quest for an Urtext was strong, the ›composer's intentions‹ loomed large over musical editions. The goal was, at various times, an ›original‹ version, a best version, a final version, an autograph version, a version conforming in its editorial markup to all other works by the same figure. Looking back on a century-and-a-half of editing music for scholarly use, we have learned that political, economic, and cultural matters all bear on the interpretation and adjudication of specific editorial issues. Tastes change. We believe that all notions of a ›best‹ edition in our own time are likely to be continually modified by future generations. Therefore, we believe, there is a clear ›philosophical‹ advantage in encoding music in such a way that it can be re-edited repeatedly, as new material and changing values warrant. This attitude may help to explain why we have been slow to provide user tools for the use of our data. User tools are still highly dependent to rapidly changing technology, but data is durable.

As our data holdings grow, so do the most practical means for distributing it. Currently the *MuseData* corpora contain more than 1,000 works, chiefly from the 18th century.[4]

- Bach: most of the instrumental music, the masses, motets, and oratorios and many of the cantatas.
- Corelli: the sonatas and concerti grossi of the six published volumes.
- Beethoven: all of the symphonies, the violin concerto and several quartets.
- Handel: the printed opuses of instrumental music (3, 4, 6, 7); eleven operas and oratorios including *Messiah* (nine versions).
- Haydn: the later symphonies and most of the quartets.
- Mozart: the later symphonies and some chamber music.
- Telemann: c. 150 cantatas, several dozen instrumental pieces, and two major works.
- Vivaldi: six printed opuses (1–4, 8, 10) of string and wind music; the oratorio *Juditha triumphans*.

Each of these corpora has been put to practical use for somewhat different purposes and constituencies. I cite just three of them here:

- Handel: The operas and oratorios are all newly edited. All have been performed in public (by Philharmonia Baroque Orchestra and Chorus, for the Göttingen International Handel Festival, and for other productions in Freiburg im Breisgau and in Drottningholm) from ›Partituren‹, short scores, and parts produced from the *MuseData* databases. Most of the works have been recorded. *Messiah* was recorded with sufficient variant movements to enable users to program nine historical performances of the work on their CD-players.
- Telemann: In the late 1980s we collaborated with the Telemann Zentrum in Magdeburg. Although we discontinued our own work on this project after reunification, the Zentrum has been well able to carry on, and Carsten Lange reports the additional encoding of more then 500 works (in a different but well-documented code, DARMS).
- Vivaldi: Our Vivaldi scores have formed the basis of three volumes of concerto editions from Dover Publications, Inc.[17] and more are in preparation. Dover is an unusual

4 For an up-to-date inventory, see http://www.musedata.org.

publisher because it does not provide performing materials. Performance materials for the Dover (score) editions and full scores of other works are available on our website.[6]

Challenges

What are the biggest challenges facing those who are preparing electronic corpora? In a word, they are the challenges, which we as musicologists are not trained to deal with. They are issues of (1) distribution, (2) copyright, and (3) user access.

(1) Distribution

Distribution becomes a more complicated subject every day. It is complicated by the multiplicity of ways in which electronic data can be conveyed and the many ways it can be used. It is complicated by the fact that data can be easily modified (by accident or intention). This multiplicity supplants a unidirectional, unambiguous system of preparation and production which has been in use for 150 years. That is, we traditionally imagine that the process of ›creating‹ and distributing a musical work begins with a composer's idea, proceeds to a manuscript draft, the typesetting of the work, and, if possible, the recording of the work. What are distributed are the typeset score/parts and one or more recorded performances of the work. In the virtual world of computer data, the work may originate electronically. It can be recycled over and over through various electronic processes to create a score, or a MIDI file, or a sound assemblage layer by layer. The composition may be algorithmic and never produce quite the same instantiation from one rendering to the next. What exactly is most advantageously distributed?

A radical reduction in preparation time results from the electronic preparation of musical scores. In place the of Bach Gesellschaft ›half-century‹ model of production, it is now possible to encode large repertories in a decade of less, provided that adequate staff resources are available. (An electronic version of the *Neue Mozart Ausgabe* current in preparation aims for completion by 2006, a mere three years after its gestation.) Critical notes can be integrated with the edition itself (as pop-up texts, as conventional notes given in a separate file, or – as in the case of the C. M. von Weber edition – reconstructable by the user from facsimiles of available manuscripts and/or early prints shown in parallel with the modern score).

The conventional paradigm of series, volumes, and pages is more or less irrelevant in electronic publishing, except insofar as earlier exemplars from the print era are cited. Yet performers of art music still learn and play from the printed page. A beautifully printed score is still a pleasure to read. A palpable artifact is still indispensable to the discussion of the de-

5 *The Four Seasons* and other concertos from *Il cimento dell'armonia e dell'invenzione*, Op. 8 (New York, 1995), *L'estro armonico*, Op. 3 (New York, 1999); Six Flute Concertos, Op. 10, with Related Concertos for Other Wind Instruments (New York, 2002).

6 The Center's homepage is at http://aaa.ccarh.org. Scores ready for downloading are at http://scores.ccarh.org. Data in several formats for downloading is at http://www.musedata.org. The Themefinder website, for searching incipits by use of symbolic codes for musical features is at http://www.themefinder.org. The Haydn-Mozart quartet quiz (style discrimination) is at http://qq.themefinder.org.

tails of a ›Notentext‹. It seems unlikely that printed music bound in books and fascicles will disappear. The expenses of conventional music-publishing have become prohibitively expensive for many purses, however, and electronic distribution has much to recommend it.

Even though our goal is to serve a broad public with minimal barriers (especially financial ones) to access, we have learned that we have to set limits. We currently permit registered users to download a maximum of 100 files a day. Quantifying music is an uncertain proposition in the best of times. One hundred files may amount to only two symphonies or half an act of an opera; on the other hand, it could be greater than all the files needed for one book of Bach's *Well-Tempered Clavier*. (Users undertaking serious projects with large amounts of data are encouraged to request the data on a CD-ROM, since downloading large quantities of material can be also process.) We have found that without some kind of control, our server is raided by commercial robots and attacked at random by hackers. As users of other people's data, we are as likely as others to find well-intentioned restrictions annoying and sometimes disabling. Finding a happy medium between control and ease of access is a problem which will not be solved by musicologists. Yet we encourage musicologists to take an interest in these issues, for they affect the future of the discipline. Musicologists' insights are as welcome and as likely to be valid as those of other professionals.

(2) Copyright

Copyright is a more fundamental and a more complicated issue. Of necessity, CCARH has been studying music-copyright issues almost since its formation. Regrettably, we principally encode out-of-copyright editions, such as the Bach Gesellschaft, because no permission has been forthcoming (despite extensive discussion) to do otherwise. In some cases, however, we commission new editions or collaborate with other researchers on the preparation of new editions. (Many of our Handel and Vivaldi editions are new ones; most of the Telemann editions were collaborations made at a time when the German editors had little access to modern tools.) As developers and users of electronic data, we wonder, though, whether a generation of work on current paper ›Gesamtausgaben‹ will be ignored by future generations – not because of any failure of quality but because of the continuing commitment to the 1850 paradigm of a physical, as opposed to a virtual, publication. Although the preparation of almost all new editions of music involves the use of a computer, some methods of preparation are not all suited to multiple uses and recycled data. Some projects which have good potential to be transformed into virtual electronic editions lack leadership knowledgeable about the ways and means of migration to a newer model. Planning at the outset is thus very important. The fear of misuse causes some music publishers routinely to discard such computer files used in score preparation. This practice potentially increases the cost of preparation of future editions dramatically and unnecessarily. The underlying questions are similar to those for distribution, however. How much access is desirable? How much control is necessary?

In European art music the differences in copyright provisions between countries, not to mention general differences between Europe and the US, perplex the most knowledgeable of legal scholars. The table below calls attention to a few outstanding differences between US and European copyright provisions related to music.

Europe	US
United Kingdom: Recognizes graphical-image copyright to be separate from content copyright.	No recognition of separate rights for graphical image (e.g., a page of music with a particular layout, size, fonts, etc.).
United Kingdom: Protects assiduous labor involved in editing of materials.	No protection of »sweat of the brow« (e.g., in careful proofing of a musical text resulting in minor changes to the content).
France: No right of fair use.	Permits quotation of small portions of a protected work (score, recording, etc.) for academic and scientific purposes.
Germany: Recording acceptable as primary instantiation of a musical work.	Primary instantiation of a musical work considered to be a score.

We have participated in two symposia on the legal issues surrounding virtual scores; one at Columbia University's Kernochan Center for Law and the Arts, May 2003; the second as part of the »International Symposium on Music Information Retrieval« (ISMIR), Barcelona, October 2004. A third dedicated meeting is currently under preparation at Stanford University. These efforts are meager relative to the degree of complexity embedded in the underlying questions.

Methods of financial support for the making of such critical editions appear to play a substantial role in the current interpretation of rights and permitted uses of materials. The conventions which were in effect for the last half of the 20th century can easily be viewed as having had strong cultural and political motivations rooted partly in the Cold War. As German reunification has caused changes in these identities, so too it has disturbed the financial underpinnings of ›Gesamtausgaben‹. What seems to be needed now is a new cultural framework for international collaborative enterprises. Ours is a small profession and international collaboration continues to be highly desirable. We must share ideas and learn from one another if we are to go forward.

(3) User Access to and Interfaces for Musical Data

We at CCARH have never regarded the development of user tools as our main goal, but we recognize their essential importance. We have sometimes been dismayed to see how weak has been the acceptance of computer technology by musicologists, and how timid musicologists are when confronted by unfamiliar technology. This can be attributed to two phenomena:

1. Students in the humanities are not encouraged to learn anything about how computers work. While this is fine by itself, some teaching of negative values discourages the curious and the energetic. Students are taught to prefer ›off the shelf‹ applications with ›glossy‹ graphical user-interfaces. Consequently, the possibilities for musicologists to investigate new kinds of applications and to refine and extend existing research methods to reflect their own interests and values remain invisible. Many potentials go unrealized. Intellectual prejudice is not a technical problem; it is a social one that ill befits the academic community.

2. While musicologists may studiously ignore nascent opportunities, hundreds of librarians, electronic publishers, computer scientists, software writers, mathematicians, cogni-

tive scientists, audio engineers, and a host of other professionals are promoting the field of ›music research‹ without the knowledge that the discipline of musicology exists and without benefit of the insights that musicologists could bring to their enterprises – including the preparation of musical editions. The best students in our seminars on music-representation and music-query come increasingly from computer science, engineering, and cognitive science, with relatively slight involvement of musicology students. Technically oriented students often have a solid training in musical performance and frequently have a good grasp of music theory, but few see the value in distinguishing between a primary source and a secondary one.

There are two paths by which to ameliorate this bifurcation between technology and musicology:

1. The first is to develop friendlier user-interfaces for software that is relevant to the tasks that musicologists are likely to want to perform. Some examples of efforts to do this can be found in the Wagner-oriented *JRing* browser of Andreas Kornstädt and the harmonic-analysis visualizations[7] (›keyscapes‹) of Craig Stuart Sapp. Special skills are required to develop such implementations, but the best results will come, I believe, from collaborative projects involving musicologists, software designers, and perhaps engineers.

2. The other is to initiate collaborations between historical and systematic musicologists; for systematic musicologists may be best able to bridge the intellectual divide between technologists and historical musicologists. In our own case, the most rigorous and constant use of our data has come, somewhat surprisingly, from experimental psychologists. Music-psychologists, many of whom are involving in extending the reach of music theory into cognitive domains, bring rigorous statistical training together with a solid grounding in conventional repertories to their search for better understanding how human beings relate to music. In a sense they are practicing systematic musicology under another label.

Towards the Future

In scholarship throughout the humanistic disciplines, the additional value of the searchable and modifiable computer materials that are by-products of electronic editions has been widely recognized – except in music. The existence of growing electronic corpora has rekindled interest in systematic musicology in many interdisciplinary environments. From our perspective there is ample opportunity for cross-fertilization between historical and systematic musicology (both broadly defined).

Historical musicologists cannot maximize the value of their electronic editions, however carefully they may be edited, without using some of the new capabilities (e. g., the open-source music-analysis tools developed by David Huron, Craig Sapp, and others as the *Humdrum Toolkit* that systematic musicologists have been using for more than ten years).

7 The *JRing* software is discussed in: Andreas Kornstädt, »SCORE-to-Humdrum: A Graphical Environment for Musicological Analysis«, in: *Computing in Musicology* 10 (1995 / 1996), p. 105 – 122. A basic explanation of keyscapes is given at http://ccrma.stanford.edu/~craig/keyscape/.

Other tools of future value to musicologists include the intensive efforts of the present to develop music-repertory-search capabilities which are cognitively informed (e. g., the recent work of Daniel Müllensiefen and Klaus Frieler, Goldsmiths College, University of London) and music-analysis tools which give users access to multiple cognitive and perceptual levels of the work instantaneously (e. g., Craig Stuart Sapp at Royal Holloway College, University of London, and Anja Fleischer Volk, Dept. of Information and Computing Sciences, University of Utrecht).

Our hope is that the bridges that are to be built between the traditionally separate fields of historical and systematic musicology will enhance the value of musical data and the editions it is used to create not only for notation and source-control but also for sound and archiving applications. Its effort can be extended for analysis and for methods of representing analytical results which improve their comprehensibility to non-specialists. Such possible developments promise to open up new application areas in pedagogy and to encourage broader public appreciation of important repertories. This is an ambitious agenda, but if we can achieve only a portion of it, we will have significantly increased the value of the arduous labor that underlies critical editions and will help to preserve the musical legacies we have inherited for future generations to understand and enjoy.

Therese Muxeneder (Wien)

Das offene Archiv

Philologie und virtuelle Sammlung am Beispiel des Nachlasses von Arnold Schönberg

Kulturelle Identität ist in Quellen eingeschrieben, deren historische Bedeutung und Wertigkeit durch die Bewahrung in Archiven dokumentiert wird. Die Errichtung von Archiven als Sammelstätten unikater Objekte, welche als Teil einer Überlieferungsgeschichte fungieren, reflektiert die (kultur-)historischen und -soziologischen Präferenzen bestimmter Epochen und Zeitströmungen. Charisma und Verführungskraft des Archivs können auf die Authentizität und Einzigartigkeit der dort aufbewahrten medialen Träger zurückgeführt werden. Die Diskrepanz des Archivs basiert jedoch auf einer ideell unendlichen Zeitperspektive als Auftrag und Legitimation bei gleichzeitiger medialer Endlichkeit. Die Gratwanderung besteht darin, sowohl den konservatorischen Auftrag wahrzunehmen als auch die Quellen in einer möglichst effizienten Form zugänglich zu machen. Die Anforderung der wissenschaftlichen Arbeit an einem Nachlass schließt präventive Konservierung und inhaltliche Erschließung gleichermaßen ein. Aufgrund der avancierten technischen Möglichkeiten im digitalen Zeitalter ist es nunmehr möglich, beide Interes-