Data

Data Bases of Text: Approaches to Organization and Access

In previous years we have provided detailed listings of a large number of data base projects, most of which are still running. Readers may consult the comprehensive index in the 1989 edition of *Computing in Musicology* to locate information about these projects. The need for a comprehensive bibliographical list of data base projects maintained by both individuals and groups has been recognized by the International Musicological Society's Study Group on Musical Data. In this issue we turn the focus temporarily to issues of organization, dissemination, and access relevant to all data bases. Projects that illustrate the issues discussed are cited. Readers are urged to continue reporting their work for inclusion in future issues.

Data bases form a central pool of information for an increasing number of scholarly collaborations. Working groups attempting to establish guidelines for new data bases confront many issues in common, irrespective of whether they are international collaborations or desktop projects. Such questions as (1) whether the data is raw or interpreted, (2) whether it is complete ("fulltext") or selected ("structured"), (3) whether it is intended to generate a printed result and/or to support online searching, and (4) whether it is fixed and final or a bank of information that is periodically updated differentiate one kind of data base from another. In addition to the general issues of organization that all data bases face, the nature of the topic may provide opportunities and raise issues of its own. We have attempted to underscore these considerations in the following discussion.

Fulltext Data Bases

A fulltext data base is a machine transcription, or encoding, or a complete work or works. Shakespeare's plays offer an illustrative example. Encodings of different editions are currently available from at least four sources—a commercial software firm (CMC Research, Inc., 7150 Southwest Hampton, Suite C-120, Portland, OR 97223), an academic press (Oxford University Press), an academic computing service (the Oxford Text Archives), and a hardware manufacturer (NeXT, Inc.). The range of choices available to those seeking an online Shakespeare may be suggestive of choices that will someday be available for musical repertories and treatises. For scholars, the question right now is often, "Is this possible?" In a few years it could well be, "Which source do I want to use?"

Fulltext data bases can be created by keyboarding or by optical scanning. The latter technology is more labor-intensive than many project planners realize, since data verification can be more arduous than for ordinary typescript. Scanning errors tend to be very subtle ones that are neither easily recognized by the naked eye nor anticipated on the basis of experience with text created by typists. Optical scanning produces an electronic facsimile. [Discussion of the technical aspects of optical scanning occurs on pp. 36-7. Keyboarding provides many choices for encoding.

Technical facility supports much that existing copyright law does not allow, and this gap is the most serious issue confronting scholars wishing to develop fulltext data bases. In the case of fulltext data bases, recent scholarship that already exists in published form is less likely to be made available for electronic distribution than works from the nineteenth and earlier centuries, even when the new work has been created from an electronic manuscript. Since this problem is broadly relevant to electronic publishing, it is more likely to be resolved in the commercial sphere than in the academic one. Yet scholars need to take cognizance of it.

Fulltext data bases exist in the first instance without support software. When such data bases are distributed commercially, they are often bundled with software that permits structured searches of specified kinds. Despite the completeness of the text, users may not be able to make up their own questions or to browse through or print large portions of the text. Such restrictions, when imposed by the software, protect the developer's investment of time in creating and correcting the text but may limit its potential uses. To put these considerations in perspective, we are profiling below two projects directed toward similar goals of encoding of medieval music treatises. These projects are called *THEMA* and the *Thesaurus Musicarum Latinarum*.

THEMA

THEMA is a textual data base of more than two megabytes developed by a single individual, Sandra Pinegar, in the context of research for a doctoral dissertation. It comprises music-theoretical texts in Latin of the thirteenth and fourteenth centuries. Many of the treatises are anonymous and most are not exactly dated. It is illustrative of efforts to capture raw data electronically.

Texts of this period are highly abbreviated and a system of encoding abbreviations by two-digit "tags" is a unique part of *THEMA*, which is designed to reflect the actual appearance of the text as well as to convey the text itself. This encoding system allows determination of the density of abbreviation in a source and documentation of changes of abbreviations from one gathering to another. It can aid in tracing some scribal errors to misreadings of abbreviated words found in other exemplars. Even changes from gathering to gathering can be examined. Approximately three dozen works by such figures (or hands) as Walter Odington, Magister Franciscus, Jerome of Moravia, and Anonymous IV have so far been encoded.

THEMA is readily transportable for use in a variety of word processing and concordance programs and for use with a variety of hardware configurations. It has been developed in an MS-DOS environment. Because it forms an integral part of dissertation research, methods of distributing the data to other scholars have not been a primary concern. In principle, however, the emphasis is on creating an electronic transcription that, once verified, will remain fixed. Enquiries about *THEMA* may be addressed to Sandra Pinegar, Music Department, Dodge Hall 703, Columbia University, New York, NY 10027.

THESAURUS MUSICARUM LATINARUM

The *Thesaurus Musicarum Latinarum*, a collaborative project led by Thomas J. Mathiesen at Indiana University, is intended to include fulltext encodings of all music theory treatises in Latin from the sixth through the mid-sixteenth centuries.

Previously encoded materials in Greek and Latin have been amalgamated with optically scanned nineteenth-century editions by Coussemaker and Gerbert to produce a first phase of data capture. It is hoped that the project will eventually include the series *Corpus Scriptorum de Musica*, *Divitiae musicae artis*, the Colorado College Critical Texts, and unedited manuscript sources. This project is illustrative of the use of edited and interpreted data, since it incorporates modern transcriptions and requires regularizations of presentation and typography.

In the *TML* spellings are normalized, diacriticals are omitted, abbreviations are expanded, and proper nouns are capitalized. Musical symbols will be entered according to a table of alphanumeric codes that include shapes of notes and ligatures, coloration, mensuration and proportion signs, and so forth. In contrast to most of the encoding schemes for early notation discussed earlier [pp. 23-35], this approach is a text-oriented one stressing the identity of each object but not expressing the relationship between objects. It is highly specific, however, distinguishing, for example, between a square flat sign ("sqb") and a rounded one ("rob").

Eventual distribution through a mainframe-based online listserver at Indiana University is the current goal. The intended uses are to facilitate studies of terminology and the preparation of concordances and new critical editions, which in turn can become part of the data base. The *TML*'s address is Dept. of Music, School of Music, Indiana University, Bloomington, IN 47405 (MATHIESE@IUBACS.BITNET or MATHIESE@ UCS.INDIANA.EDU).

Structured Data Bases

The material in a structured data base has been selected and organized in some way to facilitate its use. Structured data bases include those created with commercial relational data base software, such as dBase III or ORACLE. Structured data bases suit many kinds of subject matter. In particular we have reported a large number of projects concerned with cataloguing instruments by type or location and with cataloguing repertories. Several data bases devoted to repertorial histories of opera (the Verdi project at New York University, the Puccini project at the Technical University in Berlin, work at the Wagner Archives in Munich) are also under development. Generally these data bases are constructed and maintained by one or a few individuals. The largest and best known relational data bases are bibliographical ones developed and maintained by library consortia; we reported extensively on three of these—RISM, RLIN, and OCLC—in our 1988 issue [pp. 11-32]. Since structured data bases require user selection of material, the copyright inherently lies with the person who constructs the data base. The central questions concerning structured data bases are what to include, how to arrange and record the information, and how to disseminate the data. It is really methods of dissemination that most readily distinguish different categories of data bases. For large bibliographical data bases, the current direction is toward distribution on CD-ROM. Some representative examples from the field of music are the following:

- MUSIC CATALOGUE OF THE NETHERLANDS (MCN). The *Music Catalogue of the Netherlands* sells its catalogue of 200,000 comprehensive title descriptions of printed music, in principle allowing every detail to be retrieved on a CD-ROM with system messages for IBM-PC compatible machines in English and Dutch. Enquiries may be sent to MCN MUSICROM, Postbus 119, 1200 AC Hilversum, The Netherlands.

- OCLC. The Online Computer Library Center, Inc. is offering two sets of bibliographical records on CD-ROM. One set facilitates cataloguing while the other supports reference searching. Both are part of the *CAT CD450* system, a desktop reference library of bibliographical data bases. The *Music Cataloguing Collection* contains records for nearly one million sound recordings and music scores in one alphabetical file on two discs. The sound recordings data base, called Music Library, is also separately available in a format that permits queries for individual works and supports customization of records. It is based on more than 400,000 citations from US libraries. The contact address is OCLC, 6565 Frantz Road, Dublin, OH 43017-0702.

- OLIS. *OLIS*, an orchestral repertory data base listing more than 4000 titles, is built in *Advanced Revelation* and operates on IBM PC's. It includes information about artists, contracts, concert attendance, and premières. Developed originally for in-house use by the American Symphony Orchestra League, OLIS has been made available as a bibliographical resource. Enquiries may be sent to OLIS, 777-14th St. NW, Washington, DC 20005 (202-628-0099).

■ RILM. A CD-ROM containing the entire contents of *RILM* from 1970 through 1984 is scheduled for release in the autumn of 1990 by the National Information Services Corp. Seventy thousand abstracts from more than 300 music journals will reside on the disk, which is available by subscription. Periodic updates and annual reissues are planned. Information is available from Fred Durr, 6, Wyman Tower, 3100 St. Paul St., Baltimore, MD 21218 (301-243-0797).

The distribution of individually maintained data bases has been inhibited in recent years by lack of compatibility between commercial data base programs and by the absence of recognized conduits for distribution. The first obstacle has been resolved by the software industry. The second is being redressed in the sciences and in business applications, as large clusters of data on related topics but in diverse formats are assembled in more general formats, with appropriate search tools, on CD-ROM's.

Multi-use and Unpublished Data Bases

Publishing technology is increasingly supportive of data distribution in multiple formats. The content consists of complete texts or selected information. A hardcopy publication in the sciences may also be available (although at considerable expense) as a set of graphic images of the pages on a CD-ROM. These images (constituting an electronic facsimile) are viewable but not machine-searchable. The two sets of information remain identical.

Some authors and publishers are also supporting dual modes of publication consisting of a hardcopy product in fixed form and a machine-readable source that is periodically updated. One example is provided by Charles Mould's dBaseIII+ information bank on keyboard instrument makers. Mould's immediate objective was to create copy for the third edition of D. H. Boalch's *Makers of the Harpsichord and Clavichord, 1440-1840.* The conventional book is scheduled for publication in 1991 by Oxford University Press, but scholars will also be able to gain access to a periodically updated data base, which currently lists more than 350 instruments not described in the second edition.

In other cases there may be hardcopy and electronic complements to the same set of information. This is the case with Robert M. Keller's *Dance Figures Index: American Country Dances, 1730-1810.* The figures of 2738 dances have been encoded (for example, OXR = "Circle, Hands across, Right and Left") and are listed by title, by figure, and by page location in a book of 120 pages (ISBN 1-887984-04-3). The data itself is available in dBase or IBM DOS text-file format on a companion diskette (1-

877984-05-1). Both items are available at modest cost from the Henrickson Group, PO Box 766, Sandy Hook, CT 06482.

In some cases the original purpose of the data base is to facilitate the creation of indices that are in turn designed to facilitate access to original source material, so they are concerned with information on two levels--the data immediately available and the sources on which this is based. The *Register of Musical Data in London Newspapers, 1660-1800*, which is based at Royal Holloway and Bedford New College, aims to facilitate extraction of all musical references contained in London newspapers by organizing and indexing them. It uses the ORACLE database management system with the query language SQL. Key word searching and free access to the text base are both supported. Further information is available from Rosamund McGuinness, Dept. of Music, Royal Holloway and Bedford New College, Egham Hill, Egham, Surrey TW20 0EX, UK.

Data bases designed to provide access via a personal computer to bibliographical information are proliferating in countries in which telephone access charges are not prohibitive. Such means of access are the intended mode of operation for such US projects as *Jazzbank*, a discographical data base under development by David Robinson Jr. on the ORACLE relational data base management system, and the *Union Catalog of Black Music Holdings* developed by Samuel A. Floyd Jr. with the STAR data base and information retrieval systems.

Syntagma Musicum, a bibliographical data bank of recently published musicological articles based in Turin, Italy, is unusual in that it permits users to type in messages to other users and to make information of their own available. The main entries are assembled monthly from periodicals available at the Della Corte Civic Library. Publications seeking inclusion may be sent to the Istituto di Musica Antica Pamparato, via Gioverti 75, I-10128 Turin, Italy, with a request to open a free subscription. Schematic classification follows RISM guidelines. Entries are kept as short as possible. The data bank can be reached by telephone (modem) at 011-39-59-62-75. The data bank is supported by the city government. No special software is required. User access is free.

The creation of in-house catalogues with potential for multiple methods of distribution constitutes a perennially important sphere of activity. Among newly reported projects, that of the Centro de Documentacion Musical de Andalucía (Carrera de Darro, 29, 18010 Granada, Spain) will interest those engaged in the study of Spanish music. It is concerned with the music and dance of Andalucia and extends to inventories of original musical sources, modern editions, recordings, and instruments. The collection of this information will facilitate the creation of an encyclopedia of music and dance in the province and will also provide a basis for future editions and recordings.

MusikkFUNN: A Music Network

Nothing reported in this section matches the ambitious plans for *MusikkFUNN*, a music information network intended for operation within the FUNN framework supported by the government of Norway. Fourteen FUNN centers have been designated. They will serve individuals, schools, research centers, business and industry, and the public sector. Each center will provide information services for its own geographical region. The intention is to provide facilities that, because they are of the highest quality, would be prohibitively expensive for most enterprises to maintain individually.

MusikkFUNN, as conceived by its originators at the Western Norway Research Centre, will provide access to bibliographical information [taking as its model the *Musiek Catalogus Nederland*; see p. 134], address lists for amateur and professional musicians, tools for the preparation of concert programs, and optical disk storage of musical scores. Links with a proposed National Centre of Music Technology, coordinated by Arvid Vollsnes of the University of Oslo, are also intended. An associated data base of documents relating to the life and music of Edvard Grieg [see pp. 153-4] is currently in preparation at Sogndal College of Education. Further particulars may be obtained from Dagfinn Bach, Project Manager, MusikkFUNN, Vestlandsforsking, Fjørevegen 17, P.B. 163, 5801 Sogndal, Norway.

Data Bases including Musical Incipits

Data bases of musical incipits involve the encoding of musical information together with bibliographical records that provide information about the source. They can be designed for any of the kinds of dissemination mentioned above, although the possibility of corruption in direct electronic transmission poses serious problems to data integrity.

The most extensive bank of encoded musical incipits is that maintained by RISM in Frankfurt for the indexing of seventeenth- and eighteenth-century manuscripts of European music. More than 100,000 incipits and associated bibliographical data can now be retrieved by personal computer. RISM's data is designed to facilitate the creation of catalogues of sources. These have already taken the form of hardcopy books devoted to the holdings of one library and microfiches of aggregate sources. Other methods of access and dissemination may occur in the future. RISM's musical data, which was described in the 1988 *Directory*, is encoded in Plaine and Easie. A facility for screen display of musical information has recently been developed at the project's central headquarters in Frankfurt and is to be used by French and English working groups. Further information is available in *INFO RISM* No. 2 (April 1990), pp. 7-17, and from RISM-Zentralredaktion, Sophienstr. 26, D-6000 Frankfurt/M. 90 (069-70-62-31).

Another ambitious project, in terms of the quantity of information involved, is the National Tune Index series, which is concerned with providing linked title and letter-code music listing of tens of thousands of British and American works (songs, dances, ballad operas, wind band music et al.) of earlier centuries. The quantity of information is immense, and the listings are currently provided on microfiche through University Music Editions (P.O. Box 192, Fort George Stations, New York, NY 10040). A guidebook by the project's originators and directors, Kate Van Winkle Keller and Carolyn Rabson, is also available. NTI data consists of pitch and stress information and is cross-referenced by incipits given in scale degrees, incipits given in stressed-note sequence, and incipits given in interval sequence as well as by titles, tune names, first lines, and so forth.

Many individual data bases of musical incipit information are developed to support the creation of thematic indices. In relation to catalogues of repertory, two of the biggest undertakings—Harry Lincoln's *Italian Madrigal Indexes* and Jan LaRue's *Thematic Identifier* for his *Catalogue of 18th-Century Symphonies*—have recently come to fruition. Lincoln's catalogue, published by Yale University Press in 1989 is now the model for a sixteenth-century motet index. It provides the music itself and various indices, such as an intervallic sequence index, in the same volume. LaRue's work, planned for three volumes, provides a letter-code thematic locator in the first volume (Indiana University Press, 1989); musical material will follow in ensuing volumes.

Musical Data

***MAPPET* from Essen University**

The most significant release of musical data suited to academic use over the past year has been of folk materials encoded over many years at Essen University in Germany. These materials, consisting mainly of German and Chinese songs, have all been encoded in *ESAC*, the Essen Associative Code. This is an alphanumeric scheme for the representation of pitch and duration. Lyrics are not included.

The *LIED* and *BALL* data bases of German folk materials contain approximately 6000 melodies. The *ETHNO* data base contains information on approximately 4000 works from many cultures available on sound recordings. The *LIAO* data base combines features of the other data bases in that it contains information on 1500 recorded Chinese folk songs, and it contains encoded melodies for almost 800 of these. *ETHNOBIB* and *EDVLIT* are bibliographical data bases citing books and articles on ethnomusicology (537 items) in the first case and computer applications in music (1113 items) in the second. DIAS contains short descriptions of 1573 slides related to music in the Peoples' Republic of China. *ICTM* lists 47 projects (through January 1990) of the International Council on Traditional Music involving computers. The data bases are available in AskSam and ASCII format.

MAPPET is a package of support software for using and in fact extending the musical data bases. It provides for MIDI input, editing, and storage, detects syntax errors, and supports playback and analysis (intervals, scale degrees, patterns). It also permits searches of all the data bases on a single command. Translations of ESAC code into standard MIDI (written in C for the Atari series of microcomputers), into RELAM (real-time MIDI), and into DARMS. *AskSam* queries can be accessed in English, German, French, Italian, and Swedish. The manual is available in German and in Peter Cooke's English translation.

Data and software are available by license at minimal cost from Prof. Dr. Helmut Schaffrath, Universität Essen, FB 4 - Musik - Postfach 4300 Essen 1, Germany. Two provisions of the license agreement are that additions and corrections by users be made available for inclusion in updated versions of the data bases and that the source of the data be acknowledged in publications.

Music Data from Passport Designs

Passport Designs, a commercial music software company, launched a new Music Data division in the spring of 1990. Its initial release consisted of twelve digitized repertories of professionally recorded, "presequenced" music. Works can be arranged and orchestrated for playback on MIDI equipment. The main emphasis is on popular repertories including jazz, country and Western music, rhythm and blues, and big band performances, but some classical works, including Bach's Brandenburg Concertos, are also available. *MIDI Hits* are available in Macintosh, Atari ST, and IBM PC formats. Further information is available from Music Data, 625 Miramontes Street, Half Moon Bay, CA 94019 (415-726-0280).

MusicWriter Musical Data Distribution System

This MusicWriter [in contrast to *The Portable Musicwriter*, a dedicated system for printing music that is listed on p. 70] is a support system for the distribution of encoded music on CD-ROM. The data is intended for distribution to music stores, where consumers can customize material for on-demand printing via commercial programs for music printing. Pieces can be auditioned before purchase. An alternative use of the data is to create a MIDI diskette for use on a home synthesizer. Amiga, Atari, IBM PC, and Macintosh formats are all supported. Further information is available from Jon Monday, MusicWriter, Inc., 21569 Mary Alice Way, Los Gatos, CA 95030 (408-353-2225).

Hyperware and Hybrid Products

Interactive CD's

Hypermedia products involve the linking of diverse kinds of information. HyperCard capabilities for the Macintosh have spawned dozens of specific applications for teaching and bibliography. The possibilities supported by HyperCard when linked with sound recordings and/or MIDI instruments have given rise to a new kind of teaching tool—the interactive compact disk (CD+I).

Robert Winter's *CD Companion to Beethoven's Ninth Symphony* was released in November 1989 by the Voyager Company. The *Companion*, which assumes no detailed knowledge of music, provides a lot of text information about Beethoven's life and times and a glossary of terms as well as commentary on the music. Musical notation and corresponding sound are coupled. The software is on computer diskettes, the music on a standard CD.

A similar idea is pursued by Warner New Media's *Audio Notes* series, in which the first offering, Mozart's *Magic Flute*, was published in March 1990. In this case the original recording and support materials are on the same CD's. These materials include digitized musical examples and commentary, plot synopses, and libretti in both English and German that move at the same speed as the music on systems that include a CD+Graphics player. *Audio Notes* products can also drive a laser videodisk player, and *The Magic Flute* (which occupies three CD's) can be synchronized with an Ingmar Bergman production on disk. Beethoven's String Quartet No. 14 was released in September. Partly because of the inclusion of thousands of digitized pictures, the product requires 6.5 megabytes of free space on a hard disk.

Both firms have a series of releases scheduled. Warner's list includes Brahms's *German Requiem*, Stravinsky's *Rite of Spring*, Berlioz's *Symphonie Fantastique*, and Beethoven's Seventh Symphony as well as jazz and popular titles. Voyager has also announced a *Rite of Spring*.

In a review in *Notes* (47/1 [1990], 91-7), Karl Miller writes that the *CD Companion* "is a guide to the power of media in the learning environment....It is most significant that this program was prepared by a musicologist who lays no claim to great facility with computers....[The] relative ease of construction [of CD+I tools] will likely stimulate the development of many similar packages."

Robert Skinner reviews laser videodisks—notably the Voyager Company's earlier *Bachdisc*, a series of performances of and commentary on the B-Minor Fugue from Part Two of *The Well-Tempered Clavier* by Juan Downey, and the University of Delaware's

Videodisk Music Series, a collection of recordings with scrolling scores and background material prepared by Fred T. Hofstetter—in *Notes* 46/1 (1989), 104-8.

Hypermedia Data Bases and Teaching Tools

Country Blues in Hypermedia

Adrian Freed at CNMAT, Berkeley, has been developing a data base of *Country Blues in Hypermedia*. It incorporates text, pictures, and sound. Each text line of every song is indexed to the appropriate point on the recording (scores are not indexed). Most of the sound material is from 78 rpms. The authoring platform consists of a Macintosh computer, HyperCard, SoundBase, and a Dyaxis hard disk recording system. The Dyaxis system (from Studer Editech, 1370 Willow Road, Menlo Park, CA 94025) facilitates the transfer of sound material to a hard disk. A set of external command modules was used to control the Dyaxis sound playback from HyperCard.

Music Cultures of the World

At the University of Southern California, Gilbert Blount, Charlotte Crockett, and William Alves are creating a multimedia text entitled *Musical Cultures of the World*. Written on a HyperText platform, *Musical Cultures* provides access to a glossary of terms, bibliographical material, a picture library stored on an interactive videodisk, and a sound library on an interactive CD.