Opuscope – Towards a Corpus-Based Music Repository

Thomas Noll, Jörg Garbers
Technische Universität Berlin
Franklinstraße 28/29
D-10587 Berlin
+49 30 314 73126
{noll,jg}@cs.tu-berlin.de

Karin Höthker, Christian Spevak
Universität Karlsruhe
Am Fasanengarten 5
D-76128 Karlsruhe
+49 721 608 4214
{hoethker,spevak}@ira.uka.de

Tillman Weyde
Universität Osnabrück
Neuer Graben/Schloß
D-49069 Osnabrück
+49 541 969 4458
tweyde@uos.de

ABSTRACT

Opuscope is an initiative targeted at sharing musical corpora and their analyses between researchers. The Opuscope repository will contain musical corpora of high quality which can be annotated with hand-made or algorithmic musical analyses. So, analytical results obtained by others can be used as a starting point for one's own investigations. Experiments performed on Opuscope corpora can easily be compared to other approaches, since an unequivocal mechanism for describing a certain corpus will be provided.

1. MOTIVATION

When developing new methods in structural music analysis or music information retrieval (MIR), the question arises whether they generalize to previously unseen pieces or styles of music, and how their results compare to related methods. In other branches of science, repositories with various experimental data sets are widely used for method validation and comparative studies (e.g. the UCI Repository of machine learning databases [1]). In structural music analysis, most research groups use their individually defined data sets. so that meaningful quantitative comparisons of results are not possible.

The voluminous collections of digital music scores available on the internet are ill-suited for the purpose: only few of them meet the quality standards required for musicological investigations (e.g. the MuseData library [7]). Moreover, they focus on providing a broad variety of pieces, but extracting corpora according to predefined musical criteria and the documentation of experimental results is beyond the scope of these collections.

Researchers in the MIR community have recently expressed the necessity to create standardized benchmark collections to evaluate retrieval algorithms [3, 6]. Opuscope aims at developing a corpus-based music repository which addresses these issues. Corpora of music pieces can be extracted to provide a reproducible test bed for the comparison of retrieval algorithms.

Another motivation for creating the Opuscope repository comes from the need to exchange complex music-analytical structures. This need arises for example when discussing analytical results in mathematical music theory, or when calculating new learning patterns from preprocessed data in machine learning. Analytical results can be documented in the repository referring to the unique musical corpus used, thereby enhancing scientific transparency and comparability. As a side effect, working material covered in Opuscope will focus on musical scores and performance data.

We believe that Opuscope could constitute an infrastructure from which many researchers and structural music research in general could benefit. Our vision is to provide a network that allows sharing musical data on different analytical levels to give a more complete picture of music – like observatories all over the world share their data to build a more complete image of outer space.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. © 2002 IRCAM - Centre Pompidou

2. ORGANIZATION

Opuscope consists of project groups working on specific corpora, a service team providing infrastructure and support, and end users who retrieve corpora for their research purposes (Figure 1).

Musical corpora are created and maintained by autonomous project groups in order to accommodate interests of researchers with different backgrounds. Each project group chooses a corpus of musical pieces according to criteria suiting their scientific needs and encodes them. The data can comprise works of a particular composer or genre, variations or improvisations on a theme, or different interpretations of a work. Examples of corpora include Bach chorales, "Träumerei" performances or "préludes non-mesurées." The project group is responsible for documenting the type of information in the corpus and ensuring that copyright is respected. When a corpus has reached a reasonable state of maturity, it is published on the Opuscope web platform.

The Opuscope web platform serves as a communication turntable for the participants. In particular, a list of project groups and a MUSITECH-based corpus browser will be provided (see next section). Opuscope users (possibly belonging to a project group) can retrieve a complete corpus, use it for their research and refer to it in their publications. Typical users might be a musicologist who returns a motivic analysis of the corpus to the repository, a research group that refers to Opuscope data in a publication, or a student who uses Opuscope data for course work.

The administrative tasks are taken care of by the Opuscope service team. These include the definition of the XML-based corpus description format and the registration of ongoing projects, so that researchers from different areas interested in the same corpus can join their efforts.

3. TECHNICAL CONCEPT

A corpus consists of a collection of musical data, analyses, and annotations, which can be edited and updated. In general, it will be a good idea to record a fixed version of the corpus before producing any annotations. While the corpus changes, the history of different versions will be registered using a version control system [2], where previous versions remain accessible and can be identified clearly, e.g. by their URL.

The corpus structure and additional annotations, e.g. analyses of the pieces, are encoded using an XML-based format. This format will be derived from MUSITECH, which is currently being developed at Osnabrück University along with a set of tools for a flexible representation of musical structure [4]. The MUSITECH format integrates representations of musical data on different structural levels. On the lowest level it comprises sound data and note data where symbolic data based on scores can be combined with and related to performance data, for instance from MIDI recordings. Structural and analytical data may be added to note information, containing voices, chords, motifs or other subsets representing musical structure and analytical annotations. Additionally, metadata can provide

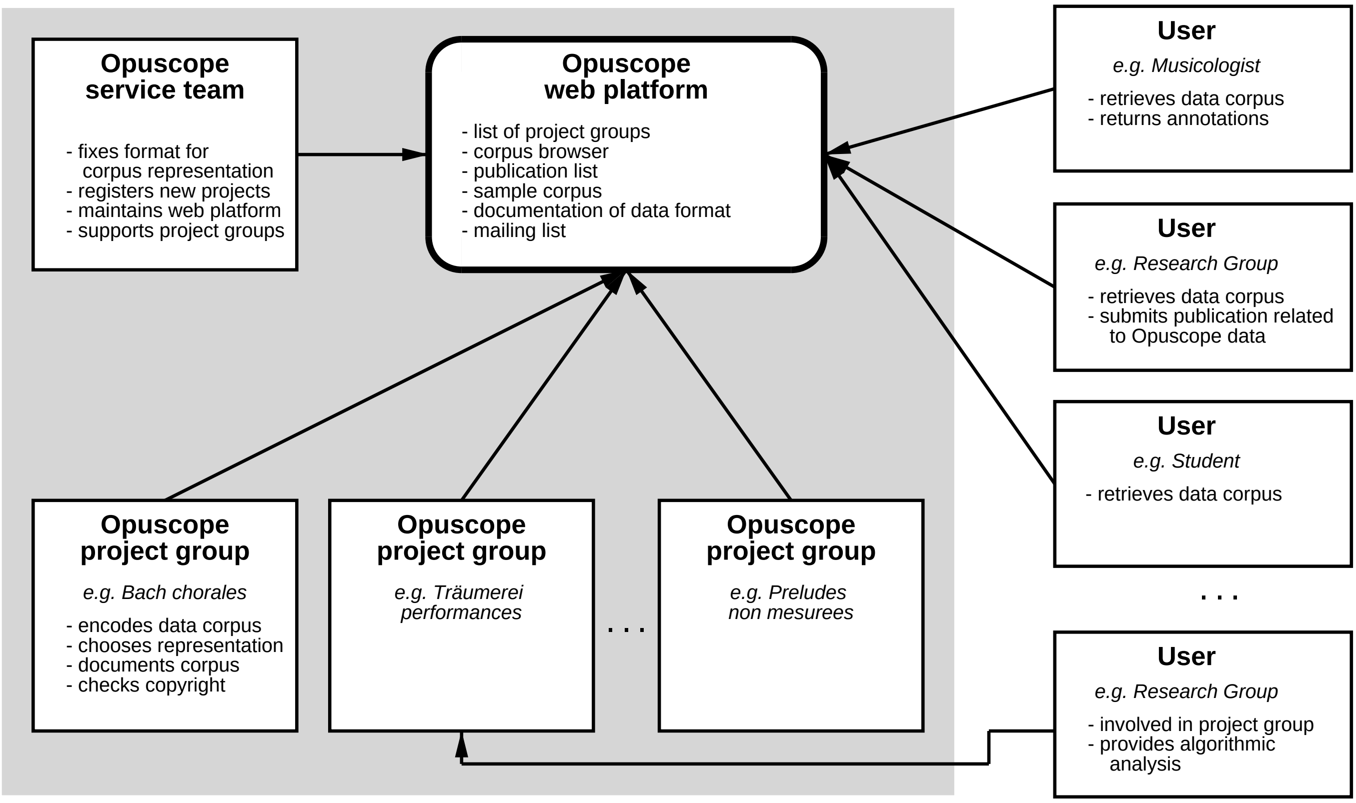

Figure 1: Organization of Opuscope

general information about the piece, the composer, and the performer as well as historical structural and other annotations wich may be useful for selecting a corpus or a corpus subset for a specific purpose.

The actual music data is stored either using existing formats such as MIDI or kern [5], or in the MUSITECH format. The idea is to have an XML file for each corpus, containing project information as well as data and/or links to external data. A web interface to browse Opuscope corpora will be integrated into the platform. In addition, the user will be able to select data on the server according to certain criteria and download only the selected files.

4. PILOT PROJECT

Currently three groups from Berlin, Karlsruhe, and Osnabrück are involved in setting up an Opuscope pilot project [8]. It is based on the infrastructure provided by the BerliOS open source software development platform, which includes a project home page, public forums, mailing lists, a project documentation manager, and a CVS repository. The idea is to implement and test a prototype of the presented concept in collaboration with potential Opuscope users.

5. REFERENCES

[1] C. L. Blake and C. J. Merz. UCI repository of machine learning databases, 1998. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

[2] CVS. concurrent versions system. `http://www.musedata.org`.

[3] J. S. Downie. Thinking about formal MIR system evaluation: Some prompting thoughts. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Plymouth, Massachusetts, Oct. 2000.

[4] M. Gieseking and T. Weyde. The MUSITECH infrastructure for internet-based interactive musical applications. In *Proceedings of the International Conference on Web Delivering of Music (WEDELMUSIC)*, Darmstadt, Germany, 2002. In press.

[5] D. Huron. Humdrum and kern: Selective feature encoding. In E. Selfridge-Field, editor, *Beyong MIDI: The Handbook of Musical Codes*, pages 375–401. MIT Press, 1997.

[6] *The Music Information Retrieval/Music Digital Library Evaluation Project White Paper Collection, Edition #1*, July 2002. `http://music-ir.org/evaluation/wp1/`.

[7] Musedata: An electronic library of classical music scores. `http://www.musedata.org`.

[8] Opuscope: a BerliOS project. `http://developer.berlios.de/projects/opuscope`.