The WITCHCRAFT Baseline Measurement and Pilot Project

Jörg Garbers, Peter van Kranenburg, Anja Volk,
Frans Wiering, Louis Peter Grijp,
Remco C. Veltkamp, Martine de Bruin

Department of Information and Computing Sciences
Utrecht University
Technical Report UU-CS-2006-064

www.cs.uu.nl
ISSN: 0924-3275

The WITCHCRAFT Baseline Measurement And Pilot Project

Jörg Garbers, Peter van Kranenburg, Anja Volk, Frans Wiering,
Louis Peter Grijp, Remco C. Veltkamp, Martine de Bruin

garbers@cs.uu.nl; petervk@cs.uu.nl; anja.volk@meertens.knaw.nl; frans.wiering@cs.uu.nl;
louis.grijp@meertens.knaw.nl; remco.veltkamp@cs.uu.nl; martine.de.bruin@meertens.knaw.nl

5 October 2006

1. Introduction

The WITCHCRAFT[1] project sets as its objective to develop a fully functional content-based retrieval system for folksong melodies stored as audio and notation, building on the best practices of Music Information Retrieval (MIR) research. Its most important goals are:

- the design of music similarity measures that are based on models of music cognition and perception;
- the creation of a search tool that is integrated in the *Nederlandse Liederenbank* and provides access for scholars and the general public to the folksong collection of *Onder de Groene Linde,* taking into account the problem of oral variation;
- to support the testing of hypotheses about oral transmission of folk songs;
- to design a music information retrieval system that is suitable for other large music collections.

The participants in the WITCHCRAFT project are:

- Department of Computer and Information Sciences (ICS), Utrecht University, in particular the Electronic Document Technology and Multimedia and Geometry groups.
- Meertens Instituut (MI), the cultural heritage institution that hosts the *Nederlandse Liederenbank* and *Onder de Groene Linde*.
- Theater Instituut Nederland (TIN), the owner of a large collection of written and recorded popular songs.

This document describes two deliverables of the project:

- Baseline measurement. Section 2 contains an abbreviated version of the baseline measurement (*nulmeting*) that has been carried out at the MI and TIN.
- Pilot project. Section 3 explains the relationship of the pilot project to the goals of the WITCHCRAFT project, and describes the pilot and its expected results.

2. Baseline measurement

The purpose of the baseline measurement is to describe the current situation at the participating cultural heritage institutes with respect to their digital resources and the need for improving access to these resources. The holdings of the MI are presented in two successive sections: section 2.1 is about the *Nederlandse Liederenbank*, a database of metadata—but not the content—of Dutch songs, and section 2.2 is about the collections of recorded and/or transcribed folksong. Section 2.3 then describes the collection of the TIN. Finally, section 2.4 describes the principal need of the *Nederlandse Liederenbank*, namely for searching the actual musical content, and the requirements for a musical search engine that satisfies this need.

2.1 Nederlandse Liederenbank

The *Nederlandse Liederenbank* (Dutch Song Database, abbreviated NLB) is a relational database that contains descriptive and semantic metadata information about c. 130.000 Dutch songs from the Middle Ages to the present day. The NLB is maintained by the MI. The database, which was created in Filemaker Pro, is accessible from the intranet of the MI. A version that is accessible to a broad public via the Internet is currently being prepared in MySQL. Several staff members spend a considerable amount of their working time on development, maintenance and data entry for the NLB.

The NLB consists of a number of sub-databases that were created in different projects and with different goals in mind. As a consequence, each of these employs a somewhat different set of metadata. None of the sub-

[1] WITCHCRAFT stands for 'What Is Topical in Cultural Heritage: Content-based Retrieval Among Folksong Tunes.'

databases provides content-based access to the melodies. For all songs that are described in the NLB a source is available at the MI, either in the library, or as a photocopy, and/or as an audio recording.

The most important sub-databases of the NLB are:

- *Bank 16e-17e eeuw* (c. 40.000 items): describes all known songs in Dutch from before 1600 and a selection of songs from the 17th century.
- *Bank Liedbladen* (c. 16.000 items): describes the broadsides from the Wouters-Moormann collection. The collection itself was digitized for *Het geheugen van Nederland* (http://www.geheugenvannederland.nl/straatliederen).
- *Veldwerkopnamen* (c. 11.000 items): describes the field recordings from 4 different collections, the most important of which is *Onder de Groene Linde*, and references to songs that are mentioned in correspondence about the recordings.
- *Bank Nederlands Volkslied Archief* (c. 63.000 items). The *Nederlands Volkslied Archief* (Dutch Folksong Archive) is a card file database that was maintained at the MI, describing songs from written sources. It was started in the 1950s and digitized around 2000.

2.2 Folksong collections at the Meertens Instituut

The database *Veldwerkopnamen* contains references to recordings of songs in four different collections:

- *Onder de Groene Linde* (7225 items, abbreviated OGL);
- *Dames Dings* (576 items), created in the 1990s by Ton Dekker & Henk Kuijer;
- *Dames Jongbloed* (122 items), created in 2004 by a group of musicology students supervised by Louis Grijp;
- *Lombok* (275 items), a multicultural collection of children's songs recorded c. 2000.

OGL, the most important collection, contains the field recordings by Will Scheepers (150 tapes) and Ate Doornbosch (380 tapes), made in various parts of the Netherlands between the 1950s and the 1980s. Many recordings were broadcasted by Ate Doornbosch in his radio program *Onder de Groene Linde* (1957-1994). Listeners would often respond to the broadcasts by reporting new variants of songs.

All recordings have been digitized (pcm and mp3). Paper transcriptions of the texts and melodies of around two-thirds of the songs in OGL are available. These transcriptions have been digitized (scanned) as well. There was no consistent methodology applied in the creation of these transcriptions, for example for the interpretation of out-of-tune singing. As a consequence, the only way to interpret the meaning of certain features of the transcriptions, is to listen to the recording.

The quality of the earliest 40 tapes is very low. In the entire collection one can expect various background sounds (cuckoo clocks, talking, passing trains, bird singing, etc). Often singing is interrupted by spoken remarks. Around 400 OGL items contain only spoken text.

The total number of performers (individual singers plus groups) in OGL is 480. 350 songs have more than one individual singer. Another 276 songs are sung by a choir or an ad hoc group of singers. Around 90% of the songs are monophonic, without accompaniment. A (substantial) minority is sung by more than one singer or with accompaniment of some sort of instrument.

For the WITCHCRAFT project, a selection of the song transcriptions will be entered into the computer as a test corpus. Funding is currently being sought for a complete encoding of all transcriptions.

2.3 The music collection of the Theater Instituut Nederland

The sheet music collection in the TIN contains almost 12,000 copies of single songs and 3400 so-called *bundels,* collections of songs. To be included in the collection a song must have been performed or meant to be performed on stage in a theatre. The collection is accessible through an Adlib-catalogue. This catalogue provides metadata such as authors and song titles, but does not include any description of the song content.

Roughly 75% of the collection was published between 1890 and 1940. Some 2000 items are preserved in handwriting or typescript. Around 90% of the songs contain text and music notation. The sound collection is much smaller than the sheet music collection. There are no transcriptions of the recordings themselves, but for approximately 70% of these corresponding sheet music exists. For 25% of the sheet music a corresponding recording exists.

By the end of 2006 all sheet music items that contain only one song will be online as digital facsimiles, as will all sound recordings. The entire digitized collection will be included in *Het geheugen van Nederland* by the end of 2006. The relevant part of the catalogue will be converted to XML for this.

The role of the collection of the TIN in the WITCHCRAFT project is to test the strengths of the melody search engine in cross-repertoire searches, and, for music research, to provide the option of tracing the origins of certain folksongs in theatre songs of the past.

2.4 Requirements for a melody search engine

The most important limitation in the present resources of the MI is the inability to search the musical content of their folksongs. Title searches provide no real substitute for these. Many melodies have a number of different texts, and one text may have several melodies. The NLB has a number of fields that describe characteristics of song texts such as rhyme and accent schemas. However, these are not available for the songs in OGL and moreover search results on these data are generally not precise enough. What the song collections need is a means for searching their musical content. Such a means must be able to facilitate the following research:

- the identification of orally transmitted melodies, i.e. assigning to them a 'melodienorm', a label for connecting related melodies;
- finding 'genetic' relations between melodies; trace variants;
- the study of cognitive aspects of melody: how do we memorize and recognize melodies, and how does oral transmission of melody work.

In addition, next to for scholars, the engine should also be usable for the general public and stimulate them in exploring this aspect of Dutch cultural heritage.

The following requirements for a melody search engine have been collected from the folksong researchers of the Meertens Instituut:

- in addition to low-level musical features, it must be possible to search accented notes, cadence notes, melodic contour (and other high-level musical features to be defined);
- it must allow approximate searching;
- it must be able to align notation and audio, and the notations of different variants;
- search output can be presented as ranked list, hierarchically, or in clusters;
- interoperability with other song collections;
- it must be possible to test hypotheses about oral transmission;
- for researchers, it must be possible to annotate search output, for example assign and store a 'melodienorm'.

The creation of an encoded test corpus is an integral part of the project. Procedures and tools for these must also be developed. Some requirements for a data entry tool are:

- an intuitive encoding system for input;
- song texts can be entered with the melodies;
- notation and audio feedback;
- audio recording can be listened to;
- high-level features can be recorded as annotation;
- software and platform independent storage format.

3. Pilot project

The aim of the WITCHCRAFT pilot project is to develop and evaluate an integrated music entry and retrieval system *prototype* for folk songs.

3.1 Relation to project goals

The pilot is in fact a first iteration in the development of the system that the project will ultimately deliver. The prototype can be employed to test the quality, efficiency and usability of methods, procedures and components. The results of this evaluation will be used in the next phase of system development.

3.2 Description and expected results

The starting-point of the pilot project is the Muugle music retrieval framework that was developed at the ICS. Different query input, matching and presentation components can be plugged into this system, as well as different music databases. In the pilot project, a data entry tool for folksongs will be developed that facilitates

the encoding of folksong transcriptions in a format that can be conveniently imported into the retrieval database. This tool will be employed to create a small test corpus. Furthermore, the Muugle framework will be adapted to the MI environment and coupled to the test corpus. Several tests will be carried out on the prototype system. Below is a more detailed description of the parts of the pilot project.

1. **Data entry and data generation**
   1.1. Develop a *music notation editor* for folk songs: the WitchCraftEditor application. This application will enable cost effective encoding of folksong melodies in an extensible format (so that still-to-be-defined high-level features can be added later). It will produce both music notation and MIDI output. A number of existing tools will be integrated into the application, such as Lilypond, an open-source music-typesetting program, and Excel (to enable annotation of high-level features). In connection with this, a suitable textual data format for folksong will be chosen, developed or adapted. A convenient user interface will be developed that makes it easy to perform the different data production tasks (encoding, conversion, viewing, listening, error correction). It will be possible to consult the already digitised OGL materials from within the WitchCraftEditor.
   1.2. *Guide and evaluate* the data entry process. A test corpus will be created that consists of 140 digitized melodies from *Onder de Groene Linde*. This test corpus will as much as possible reflect the properties of the entire repertoire, in particular the occurrence of oral variation. Various MI researchers will participate in this task. Their experiences will be used to evaluate the usability of the WitchCraftEditor, to make an inventory of the problems that emerge from the data, and to list potentially interesting high-level features.
2. **Development and evaluation of music retrieval tools**
   2.1. *Adapt Muugle* for the MI. The Muugle framework and database structures will be duplicated for testing in a Mac environment, and revised where necessary. Scripts for exporting data from the WitchCraftEditor to the database will be created. Some minor changes will be made to Muugle, such as the addition of a browsing feature for search output.
   2.2. *Evaluate the melodic similarity measures* that Muugle provides on the test corpus. For selected queries, the quality of the retrieval will be determined, by comparing search output to ground truths that are provided by music researchers. This will provide insight in the shortcomings of the current algorithms and give suggestions for future algorithm improvement.
   2.3. *Explore the potential of the tools* for addressing folksong research questions such as classification, identification and tracing of melodic variants. What functionality is lacking in the current, 'manual' folksong research practice, and how suitable are the tools for supporting this research in general and research into the oral variation of folksongs in particular? A number of best practices in folk song research will be described in use cases that will serve as a basis for further algorithm and user interface development. The study of this topic will begin during the pilot project but is not expected to be finished at the time of the pilot presentation.

3.3 Deliverables

The deliverables of the WITCHRAFT pilot project are:

1. the WitchCraftEditor application, with documentation;
2. a storage format for folksong encodings, with documentation;
3. a test corpus of 140 encoded folksongs;
4. an evaluation of the data-entry process;
5. an update of the Muugle framework for the MI environment;
6. evaluation of current Muugle similarity measures on the test corpus;
7. a draft description of the potential of the system for folksong research.

The results of the pilot project and the prototype will be presented at the CATCH meeting in December 2006.