Catafolk: Cataloguing Folk Music Datasets for Comparative Musicology

Bas Cornelissen

June 2021

1 Short abstract

We present Catafolk, a project that indexes folk music datasets and serves metadata about these collections in a consistent, easy-to-use format. A beta version of Catafolk is available on bacor.github.io/catafolk. We hope this project will help diversifying computational music research and contribute to the creation of large, cross-cultural musical datasets.

2 Extended abstract

Background A renewed interest in diversity and universality of music has motivated the study of global samples of music from across the globe (Mehr et al., 2019; Savage et al., 2015). Such samples include music from many different traditions, hoping to capture as much variability as possible. But describing a musical tradition using a single song is like describing a language using a single sentence. While clearly informative, it is not designed to map the diversity and regularity *within* a tradition: from its meters and rhythms, to its scales and modes. Large collections of music from individual cultures are better positioned to capture such properties. But those collections do not add up to a global sample, let alone a representative one.

One way in which one can start bridging this divide between 'sparse' global samples and dense 'local' samples, is by combining the local datasets currently available. Creating a large cross-cultural dataset has another benefit. Many computataional cross-cultural studies have relied on the Essen Folksong Collection, because it contains large numbers of both German and Chinese folksongs. However, choosing for Essen is usually not a principled choice, but a convenient one: either other cross-cultural datasets are not readily available, or they are not known. Since many other folk music collections have been digitized since the inception of Essen over three decade ago, it seems high time for a convenient alternative containing music beyond German and Chinese folksongs.

Aims The Catafolk project takes a step in that direction. It aims to build an index of existing folk music datasets and make essential metadata available in a consistent, easy-to-use format.

Methods Catafolk generates a csv file containing metadata for all songs in a dataset. We leave the original datasets untouched, but extract and transform the metadata in Python. The csv files are then used to generate a static Gatsby website that also makes all metadata available through a web interface. The website also documents how to obtain or use the datasets, what licences apply, and includes a detailed bibliography. Taking advanteage of Gatsby's knowledge graph, one can also query the metadata in GraphQL.

Results Catafolk is available at bacor.github.io/catafolk. The project is still in an early stage, but already contains metadata for 15,507 songs from 22 datasets. The vast majority of those are symbolic transcriptions from KernScores, the Densmore collection (Shanahan & Shanahan, 2014), and the Finnish Folk Tunes collection (Eerola & Toiviainen, 2004). Catafolk's ontology contains 61 fields, spanning musical data such as title, key, tempo or tune family to metadata on the collectors, encoders or copyrights. Entries are geocoded as much as possible, and linked to Glottolog, D-Place and eHRAF. Where possible, we also link songs to scans of the books where they were originally published.

Conclusion We have presented an early version of Catafolk, an index of folk music datasets for computational research. The project is still in development, and we are actively looking for people who want to help expanding, testing and using this resource. We hope Catafolk will contribute to the adoption of more diverse datasets in computational studies and add a further spur to cross-cultural music research.