Data Collection in Music Generation Training sets
A Critical Analysis

Fabio Morreale Megha Sharma I-Chieh Wei

University of Auckland, New Zealand
University of Tokyo, Japan

How do we collect and use music data for Automated Music Generation (AMG)? Our goal is to explore common practices of data accumulation at ISMIR and identify the implicit or explicit ideologies behind these practices. The paper highlights how current data collection methods can potentially exploit the creators of the data - the musicians. This study contributes to the emerging self-critical corpus of work of the ISMIR community, reflecting on the ethical considerations and the social responsibility of our work.

Motivations

- Rapid acceleration in AMG tasks using Machine Learning techniques
- Most of the focus is on improving the quality of generated music
- Reflections on ethical implications of AMG do not address the issue of dataset population
- The potentially exploitative nature of data accumulation goes unchecked

Objectives

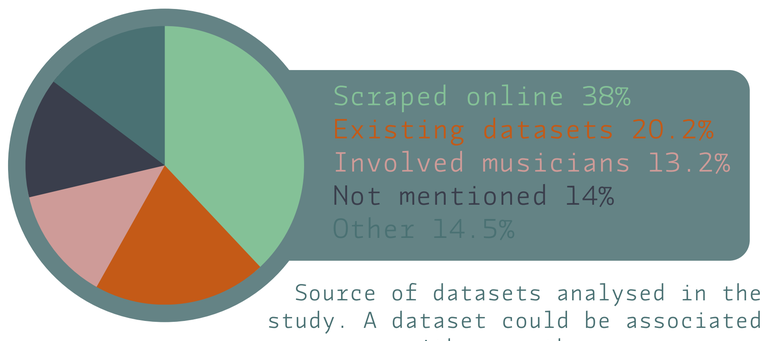
- To assess how datasets related to music generation, are used in ISMIR papers from 2013 - 2022
- To identify the involvement of musicians in said datasets
- To identify suggestions for dataset creators interested in following responsible practices in their work

Methodology

- **Identifying Papers**: we included papers presenting a new AMG model and/or introduced a new music dataset that could potentially be utilised as training material for AMG
- **Identifying Datasets**: From 121 papers selected, we identified datasets used in them, and whether these datasets were new or existing
- **Dataset Analysis**: we collected information on the 115 identified datasets: data format, source, whether it contained original content or not, involvement of musicians, and discussion of ethics in the papers.

Results

- **Dataset Creation**: New music was used in 16.5% of the datasets. Despite most datasets depending on existing or requested work from musicians, only 5 papers mention payments for this work
- **Musician awareness**: Only 3 papers explicitly mention asking permission for using a musician's music for MIR research. Problematic data collection includes lack of consent mentioned from student data.
- **Ethical Considerations**: Only 2 papers contain an ethical statement explicitly. Ethics of dataset collection are an oversight in current methods



Source of datasets analysed in the study. A dataset could be associated with more than one source.

Discussions

- **Terra Nullius**: Original music found online is considered a natural resource that is free for the taking without considering or compensating the human labour required to create music.
- **Musicians' labour** is structurally obfuscated in AMG applications to the benefit of profit and innovation
- **Musician Rights**: Musician rights are not prioritised in AMG datasets and further negligence can lead to an exploitative industry.
- **ISMIR priorities**: what and whose agendas are we implictly and explicitly following?

Suggestions for dataset creators

- **Develop one's own dataset**: shifting focus to self-made datasets, and building models trainable on smaller datasets
- **Receive consent from musicians and remunerate them**: if using music from existing musicians: consent, awareness, and remuneration are key for ethical engagement with musicians in AMG
- **Document the process of datasets development**: proper documentation should make the data collection process transparent and traceable, ensuring a check on our methodology
- **Report the intended use of the dataset**: listing potential applications can help musicians and dataset users define boundaries on the usage
- **When borrowing data, maintain the purpose of the original datasets**: where possible, consent for new applications should be necessary to avoid any unethical usage
- **Volunteer ethical considerations**: we can build our commitment towards ethical practices by voicing our ethical considerations as we build new datasets for AMG tasks