# INFRASTRUCTURES AND INTERFACES FOR DATA COLLECTION IN MUSIC INFORMATION RETRIEVAL

**Tillman Weyde and Daniel Wolff**
Department of Computing
City University London
{t.e.weyde,daniel.wolff.1}@soi.city.ac.uk

## ABSTRACT

We addressed the question of what infrastructures and interfaces would be needed to support data collection activities in Music Information Retrieval. There was general agreement that data collection is important and should to become more efficient to support MIR and related research. There was agreement that no single system could fulfill all data collection needs. However, in a quickly changing environment of web 2.0., social media, mobile web, tablet computers and smart TVs, it is very challenging for researchers to address a broad audience as needed for collection large amounts of data. The provision of reusable building blocks, a repository of experimental designs, workflows and raw data, and a supporting backend infrastructure would be extremely helpful in enabling researchers to save time and effort in designing, developing and running their experiments.

## 1. PARTICIPANTS

Jens Madsen, Sebstian Ewert, Rudolf Mayer, Amelie Anglade, Bruno Rocha, Jan Schlueter, Steve Welburn, Ashley Burgoyne, Michaela Malgas, Tillman Weyde

## 2. COURSE OF THE DISCUSSION

### 2.1 Introduction and status quo

This is a short summary of a discussion session that took place as part of the ISMIR 2012 "Demos and Late-breaking" track on 12th October 2012. The discussion started with a brief description of the CASimIR framework for online experiments, that was also presented in the demo session [1]. CASimIR is a framework for running music similarity experiments on the web and in Facebook on PCs and mobile devices.

The participants reported that they have an interest in online data collection and they currently use a variety of on-line and traditional methods. The experiments run for data collection follow a wide variety of paradigms and require individual media materials and interface elements. Several participants reported that they could currently not see a single system as supporting these different approaches. A support for different paradigms of experiments and paradigms for creation and use of data is needed, as data collection is a demanding and time intensive task, that receives relatively little reward. Therefore the prospect working more effectively and efficiently by re-using and sharing software modules is attractive.



### 2.2 Potential of an infrastructure for research

The creation of experiments and the collection of data are a scientific activity that should receive more attention with regards to the preservation of the outcomes, similar to other scientific results. The idea of building blocks that could enable the creation of an online experiment or survey with little programming effort was seen as useful, especially with repositories for preserving and sharing experimental designs and workflows.

If designs and software building blocks are reused, this could establish a culture that supports better research, because the work can focus more on scientifically relevant issues rather than on reimplementing common user interfaces.

### 2.3 Considerations in infrastructure design

A number of potential issues were brought up by the participants that can arise from reuse and sharing of software modules, such as interface components and from running experiments on the web or social media. These include technical points like the calibration of loudness levels and other parameters in listening experiments. Here, intelligent solutions could help to improve the situation, but to reach the quality which is achieveable in a physical lab will probably require great effort and probably specific hardware and may be not always possible. Similarly it is difficult to control the selection of subjects. In general, there is a trade-off between data quality and data quantity, which needs to be considered in every experimental design.

In oder to make workflow management and sharing attractive, is is necessary to create enough flexibility to accommodate very different types of experiments. It is possible that this flexibility reduces the effectiveness of the support for standard experiments. However, most participants

thought, that the potential savings of time and effort for the experimenter would still make the use of the envisaged infrastructure worthwhile.

### 2.4 Quick survey

At the end of the discussion a quick survey was held among the participants. 8 participants indicated that they are involved in experiments. 9 participants indicated that they would be interested in using an infrastructure for online experiment. 8 participants would be prepared to share their data on such a platform, provided that it is permitted legally and by their organisation. 5 participants indicated that they would be interested in programming and sharing software in the envisaged infrastructure.

## 3. CONCLUSIONS

In summary, the discussion showed a great interest by participants in the creation of an infrastructure that would support online data collection for music information retrieval and allow sharing data, software, designs and workflows for such experiments. There are concerns about the flexibility of such an infrastructure and the data quality that can be reached, but the potential benefits still seem attractive.

## 4. ACKNOWLEDGMENTS

## 5. REFERENCES

[1] D. Wolff, G. Bellec, and T. Weyde. A Music Similarity Game Prototype Using the CASimIR API. In *Proc. of the ISMIR*, 2012.