Data Management Plan

Assessment of existing data

Explanation of existing data sources used by the research project: Dig that Lick uses metadata from the J-DISC Online Jazz Discography provided by the Center for Jazz Studies at Columbia University (CU) in the City of New York (`http://jdisc.columbia.edu/`) as well as the MusicBrainz Database (`http://musicbrainz.org/`), DBpedia (`http://wiki.dbpedia.org`) and other Linked Open Data resources. We analyse audio recordings in the J-DISC collection at CU (see Letters of Commitment).
Analysis of the gaps identified between the currently available and required data for the research: The available data consists of audio recordings and discographic metadata; our interests lie with the relation between the two. In particular, we use methods from Music Information Retrieval to analyse the audio content in order to discover the information that can be garnered from automatic analysis of the recordings, and relate this to the available discographic, historical and geographic metadata and external background knowledge, in order to perform a data-driven study of the creation and spread of new musical forms.

Information on new data

Data produced or accessed by the research project: Dig that Lick will produce content-based metadata, in particular automatic transcriptions of melodic material from the audio recordings, as well as links between the various data sources used in the project. The internally used formats will consist of Sonic Visualiser project files and SQL database files and the data will be published in these formats and as Linked Open Data in a dialect of RDF.

Quality assurance of data

Procedures for quality assurance carried out on the data collected at the time of data collection, data entry, digitisation and data checking: One of the challenges of large-scale (Big Data) analysis is that it is not possible to check the correctness of automatic analysis outputs for the complete data set. Our methodology is to use existing manually created and checked transcriptions from the Jazzomat project (`http://jazzomat.hfm-weimar.de/`) in order to estimate the reliability of automatically generated data. Other automatic tests of data consistency will be performed, as well as manual checks of small random samples of the data, plus checking particular samples which give rise to the most interesting results. In addition, all provenance data will be saved and published with the data to remove any ambiguity about claims to data quality.

Backup and security of data

Data back-up procedures adopted to ensure the data and metadata are securely stored during the lifetime of the project: Data will be backed up using standard procedures in each of the partner institutions, as well as being mirrored across multiple institutions to ensure that no data will be lost during or after the project. Software will be developed using the Sound Software repository (`https://code.soundsoftware.ac.uk/`), which provides version control and redundant storage for source code.

Management and curation of data

Plans for management and curation of primary or third party data: Data will be deposited in a long-term institutional repository. QMUL provides up to 1TB of storage free of charge for its projects, using a DSPACE repository, with a guarantee of hosting the data for at least 10 years from its time of last access. We plan to publish the data as Linked Open Data using established ontologies such as the Music Ontology (`http://musicontology.com/`), in order to ensure that the data can be understood and re-used by others. The published data will contain provenance information, such as unique identifiers for source data, software details including

version numbers and parameter settings, plus details of methodology, assumptions made, and the formats and file types of the data.

Difficulties in data sharing and measures to overcome these

Obstacles to sharing data and measures to overcome these: In any project involving commercial audio recordings, it is not possible to share the recordings which we analyse. Metadata, on the other hand, is not subject to such restrictions. In order to make the outputs of *Dig that Lick* reusable and the results reproducible, we will publish links containing identifiers pointing to the analysed recordings, such as MusicBrainz IDs (MBIDs), or proprietary identifiers (e.g. YouTube URLs or Spotify identifiers). These identifiers will allow users and other researchers to access the audio recordings legally.

Consent, anonymisation and strategies to enable further re-use of data

Procedures to handle consent for data sharing for data obtained from human participants, and/or how to anonymise data, to make sure that data can be made available and accessible for future scientific research: We do not intend to gather any data directly from human participants.

Copyright and intellectual property ownership

Who will own the copyright and IPR of newly generated data: The data generated from *Dig that Lick* will be owned by the project partner who generated it. In the case of joint work, partners will own the data in equal shares, unless an agreement to the contrary is made in advance of data production. Data will be published using a Creative Commons Attribution (CC-BY) licence, in order to ensure that it can reused and extended freely.

Responsibilities

Responsibilities for data management within research teams at all partner institutions: Each partner will be responsible to comply with the data management requirements of their respective funding bodies. The Principle Investigator at each site will delegate this responsibility as appropriate to a member of their team. Data management will be a standing item on the agenda of PI meetings, to ensure that best practice is shared and followed throughout the project.