1 About This Proposal

1. This proposal describes a six-month pilot data-management project, entitled Sustainable Management of Digital Music Research Data, to run at the Centre for Digital Music (C4DM) at Queen Mary University of London from 1 October 2011 to 31 March 2012.

Objectives

2. The immediate objectives of this project are:
   (i) to develop an ongoing management plan for data produced during research at C4DM;
   (ii) to implement a system for storing and managing such data;
   (iii) to audit, catalogue, and manage existing data from C4DM using this system.
3. In addition to producing a working data management system and plan, a wider objective is to perform a pilot project toward producing a set of data management methods and recommendations which may be applicable to other research groups within the audio and music research fields in the UK. To this end, any lessons learned and systems developed during this work will be disseminated through the `SoundSoftware.ac.uk` sustainability project also managed at C4DM.

Data at the Centre for Digital Music

4. The Centre for Digital Music (C4DM) at Queen Mary University of London is one of the leading research centres in the field of audio and music technology and signal processing. C4DM makes use of a variety of data as research inputs—most obviously audio datasets—and produces a variety of types of data as research outputs. These outputs include: (i) manually annotated feature data ("reference annotations") such as expert chord and key transcriptions of existing music recordings which are used as comparative data for evaluating research work, and (ii) automatically produced annotations such as those accompanying the publication of methods for audio feature analysis. C4DM also publishes some "data sets" which are actually software services calculating data dynamically or on demand, such as those provided by the SAWA (Sonic Annotator Web Application, `www.isophonics.net/sawa/`) feature extractor and the JISC-funded LinkedBrainz project (`linkedbrainz.c4dmpresents.org`).
5. With this project, we aim to improve the efficiency and sustainability of storage and publication efforts for those data sets produced as research outputs. We will not be handling audio data (which generally can not be published, due to copyright restrictions) or data that is generated dynamically, e.g. by web services, in this proposal. We believe this restriction is appropriate given the short timescale of the present proposal and its interaction with ongoing work funded by other grants (namely the Sound Software project described below). We anticipate that outcomes from this work will prove relevant for other types of research data. In subsequent work in the Sound Software project, we intend to build on this project to cover further data types.
6. The two classes of research data to be considered here are therefore: (i) reference annotations produced within C4DM as discrete pieces of work; (ii) data accompanying research publications, such as the results of experiments, evaluation data, or representative outputs of an algorithm. The common theme is that these data originate in C4DM and are to be published externally.

Reference Annotations

7. C4DM publishes a number of reference annotations, listed at `http://isophonics.net/datasets/`, and plans to produce more in general ongoing work. Examples to date include annotations of harmony (chords and keys), rhythm (beats and bars), and structural boundaries (e.g. verse and chorus), covering the Beatles catalogue, and a number of songs from Queen, Carole King, and Michael Jackson. These annotations are used widely in the research community for evaluating algorithms for automatic analysis of audio, including in the international MIREX evaluations (Music Information Retrieval Evaluation eXchange, `www.music-ir.org/mirex`).
8. The annotations are transcribed by expert listeners attending closely to musical recordings using interactive software such as Sonic Visualiser (`www.sonicvisualiser.org`), a process that takes a significant amount of time and care. Such data typically have an associated level of confidence, based on how thoroughly the results have been checked, but depending also on the intrinsic ambiguity of the musical information being described. The data are associated with specific releases of musical recordings, although these often cannot be shared for legal reasons. (To use the data, researchers must purchase the matching recordings using the release information supplied, such as CD and track numbers, titles and artists.) Annotation data may be updated and should be versioned; the existing published data have in some cases

been updated multiple times as corrections and additions are made. This is a normal situation which must be allowed for, as human annotators can not be expected to produce perfect work.

9. Currently C4DM has only informal methods for describing and distributing reference annotations through its Isophonics web site. Files are simply copied to the web server with the descriptive information entered in the Drupal-based website. There is no policy for managing descriptive metadata, no properly maintained storage facility for such data, and no effective policy in place for backup or versioning.

Research Data

10. C4DM has produced several hundred refereed research publications in the last decade, but only a small number of associated data sets have been published. A combination of reasons could be cited for this, including failure to understand the benefits of data publishing, the lack of suitable infrastructure for managing and preserving research data, and the lack of impetus from main funding bodies (EPSRC and EU) and journals, who do not require data publication. Growing awareness of the principles of Reproducible Research applied to digital music research have led to isolated examples of publishing research data, such as the automatic chord transcriptions at `www.isophonics.net/content/automatic-annotations`, however this is the exception rather than the rule.
11. Output from automatic feature extraction algorithms is usually summarised for paper publications as a set of statistics describing the extent to which the output concurs with a set of manual annotations. This is a useful measure of progress in algorithm development, and is used for example in the MIREX series of international evaluations of music information retrieval algorithms (`www.music-ir.org/mirex`). Such statistics do not, however, tell the whole story: they fail to inform readers about which musical examples were analysed successfully, and which examples the algorithm failed on. Characterisation of successes and failures is necessary in order to gain insight about how systems can be improved. Further, a difference in mean performance of two different algorithms is less revealing than a piecewise comparison of algorithm performance, from the viewpoint of both informal and formal statistical analysis. Other reasons for data publication are to allow verification of published results and testing of reimplementations of published algorithms. Finally, algorithm outputs are needed by researchers who wish to assess others' work using different criteria or metrics than those appearing in their publications.

The Sound Software Project

12. C4DM is currently managing an EPSRC-funded Software Sustainability project, "Sustainable Software for Digital Music and Audio Research" (EPSRC grant EP/H043101/1), known as the Sound Software project with a public site at `http://soundsoftware.ac.uk/`. This project, funded until 2014, aims to provide resources, advice, and training to facilitate the development and use of software, data, and metadata in the UK audio and music research community. The focus of the Sound Software project is primarily on software resources, building the `code.soundsoftware.ac.uk` repository and facilities amongst other work, but managing research data is a complementary task which fits with the overall aims of the project.
13. The results from the project under proposal shall be disseminated via the Sound Software project and used as a basis for resources to be made available to other research groups in this field. We also plan to build on this work to cover a wider range of data types, including audio.
14. We believe that we can take advantage of the obvious affinity between the present proposal and the Sound Software project in order to enhance the sustainability of this work and of the research output from C4DM. Effectively, this proposal is for a pilot project carried out within the Centre for Digital Music on its own research data sets with its own concrete research requirements, aimed at providing lessons in good practice and a viable data management plan and system implementation which can be drawn upon by Sound Software in providing recommendations and facilities for other groups.
15. The existence of the Sound Software project also informs our selection of the types of data to focus on for this proposal. By focusing on research outputs of C4DM, we can pilot some generally applicable data management techniques in a context (data specific to this group's work) that the Sound Software project would not otherwise be directly involved in.

Users and Needs

16. The primary users for this projects are researchers (academics, post-doctoral research assistants, PhD students and MSc students) in the Centre for Digital Music (C4DM) at Queen Mary, University of London, who perform research over a range of areas including music informatics, machine listening, audio engineering and interaction. A common use-case in C4DM research is to run a newly-developed analysis algorithm on a set of audio examples and evaluate the algorithm by comparing its output with that of a

Managing Research Data - Diagrammatic Work Plan

Numbers show estimated percentage of total effort between work packages

Key: M = month, D = deliverable, WP = work package Individuals: RA | SD | All

	M1	M2	M3	M4	M5	M6
WP1 Requirements analysis	45	35				
1.1 Audit of published data	D1.1					
1.2 Audit of unpublished data		D1.2				
1.3 User requirements analysis		D1.3				
WP2 System design and implementation	45	45	45	45	45	45
2.1 Review of current platforms		D2.1				
2.2 Requirements and specification						
2.3 Implementation and integration				D2.2		
2.4 User acceptance testing						
2.5 Documentation and recommendations					D2.3	D2.4
WP3 Data cataloguing				20	35	25
3.1 Data cataloguing						D3.1,3.2
WP4 Data management plan		10	45	25	10	20
4.1 Data management plan						D4.1
WP5 Communication and Management	10	10	10	10	10	10
	D5.1					D5.2
	100	100	100	100	100	100

human annotator. Results are then compared with published results using the same input data to determine whether the newly proposed approach makes any improvement on the state of the art. This use case suggests needs that revolve around (i) the identification and procurement of suitable data for testing (typically audio input data), evaluation (reference annotations, as described aboved) and comparison data (results from state of the art algorithms applied to the same test data); and (ii) infrastructure to support the publication and dissemination of data produced at C4DM. Currently, minimal support is provided for these tasks: C4DM has no data management plan, no versioning of reference annotations, no structure for organising storage, no automatic backup facilities for research data, and only a simple web-server for making data available to other researchers. As a result, much research data is never published, and is often not even available internally, as it resides in the personal file space of researchers, and can be lost when people change institutions. Other needs of C4DM users will be determined in WP1 Requirements Analysis, described below.

17. We expect the data management methods and recommendations produced by the current pilot project to be applicable to other research groups within the audio and music research fields, and also to the wider JISC and HE communities. In order to facilitate transfer of any lessons learned and systems developed during this work, we will make use of dissemination channels provided by the `SoundSoftware.ac.uk` sustainability project managed at C4DM and the Digital Music Research Network (`www.elec.qmul.ac.uk/dmrn/`), as well as through regular JISC channels (see WP5.2 and section 3).
18. The success of the project will be recognised through the increased publication and dissemination of research data in C4DM, and the concrete recommendations for improving best practice which are fed into the Sound Software project and JISC community.

2 Workplan

19. The work of the project is divided into five work packages (WP), each allocated to either the postdoc research assistant (RA) or software developer (SD) funded under this proposal. The work plan is illustrated in the Gantt chart shown above. The required deliverables for Strand B (Call Document, paragraph 53) are fulfilled as follows: (a) D1.1-1.3; (b) D4.1; (c) D2.2; (d) D2.4; (e) D2.3; (f,g) D5.2.

WP1: Requirements analysis (Months 1–2; RA)

20. **WP1.1 Audit of data already published by C4DM**: The data published and disseminated by C4DM through existing outlets will be listed and described informally, with common qualities as well as potentially troublesome qualities for metadata description highlighted in the description.
21. **WP1.2 Audit of unpublished data**: Authors of papers published during the previous 12 months will be interviewed and an audit made of data referred to in their papers or produced during research which has not yet been made available to the public, or of data published but not yet aggregated in the usual location for C4DM datasets (at `isophonics.net`).
22. **WP1.3 Requirements analysis for access, re-use and publication of research data**: We will analyse the requirements of the authors interviewed in WP1.2, based on their use of existing research data, their knowledge of existing best practice, and any problems they have experienced in obtaining or using data from other researchers, or in making their own data available and visible to the research community. Based on WP1.2, we will determine requirements for data (types, quantity) in order to maximise the likelihood of acceptance of the proposed system by the user community.
23. **Deliverable D1.1 (M1)**: Description of data already published by C4DM
24. **Deliverable D1.2 (M2)**: Description of other data produced during recent research work but not yet published
25. **Deliverable D1.3 (M2)**: Document specifying C4DM user requirements for research data management

WP2: System design and implementation (Months 1–6; SD)

26. **WP2.1 Review of current platforms**: A rapid informal survey of existing data-management platforms and current practices, whether arising from prior JISC project outputs (e.g. MaDAM, ADMIRAL, DryadUK) or from the community at large, will be carried out and the appropriateness of each platform to the types of data being managed at C4DM will be evaluated. The Digital Curation Centre will also be consulted for advice about suitable platforms. Three possible outcomes are foreseen: (i) a suitable platform is found and chosen for the project; (ii) a platform is found that requires modification in order to make it suitable for the project; and (iii) no suitable platform is found, and the system has to be built using generic tools.
27. **WP2.2 Technical requirements gathering and system design**: The requirements process will consider the qualities of data identified during WP1, the needs of users within C4DM (WP1.3), the ability to manage backups and to publish data externally, and the hardware infrastructure available. These will be used to propose a candidate implementation. The technical requirements process will be carried out in consultation with early data management plan (DMP) development (see WP4 below) in order to promote reusability of the technical work and test the practicality of recommendations for the DMP. Given the short timescale, systems building upon existing software and deployment practices identified during WP2.1 will be preferred. Preference will also be given to open standards for storage and description, to open source software, and to software that interoperates with other systems within C4DM and with platforms being developed and deployed by the Sound Software project. It is understood that the system developed will be a pilot implementation with a goal of learning lessons for future work as much as of satisfying current requirements.
28. **WP2.3 Implementation and integration**: Development and deployment of the system to satisfy the requirements established in WP1.3 and WP2.2, using a rapid prototyping methodology. Any bespoke software developed will be published under an open source licence.
29. **WP2.4 User acceptance testing**: The system will be opened to the project RA, and handed to a pool of approximately 4 active researchers within C4DM for evaluation and to obtain feedback. This process will overlap with WP2.3 and will inform successive iterations of the software.
30. **WP2.5 Documentation and Recommendations**: The system will be documented for end users, its technical design will be described in a short report, and a set of recommendations for future implementations (including details from user feedback and experiences with WP3, below) will be compiled. Feedback from WP2.4 will be incorporated into this documentation. The recommendations will also be disseminated to stakeholders via the Sound Software project, the Digital Music Research Network and JISC Programme meetings.
31. **Deliverable D2.1 (M2)**: Summary of current platforms
32. **Deliverable D2.2 (M4)**: Data management system implementation
33. **Deliverable D2.3 (M5)**: User documentation
34. **Deliverable D2.4 (M6)**: Technical report and recommendations for future implementation

WP3: Data cataloguing (Months 4–6; RA)

35. The data listed during the audit of WP1 will be entered into the system developed in WP2, with metadata following existing schema standards where possible and in line with the recommendations being developed for the DMP in WP4. Data that are to be made publicly available will be linked through to an externally visible web resource.
36. **Deliverable D3.1 (M6)**: Catalogued data sets available to researchers in C4DM
37. **Deliverable D3.2 (M6)**: Publicly-accessible distributions of catalogued data sets

WP4: Development of data management plan (Months 2–6; RA)

38. A data management plan for C4DM will be developed, detailing issues such as the responsibilities of researchers and managers, conventions for specifying ownership and access rights, means of protection of personal data, version control and quality control.
39. **Deliverable D4.1 (M6)**: Data management plan for C4DM

WP5: Communication and Management (Months 1–6; PI, CI, RA, SD)

40. **WP5.1 Project management**: The project will be managed on a day-to-day basis by the PI (Dixon), with project meetings held weekly to assess progress and problems. This has been our practice throughout the Sound Software project and previous JISC-funded projects. The CIs (Plumbley and Cannam) will participate in the management process to ensure compatibility and continuity with the requirements of the Sound Software project from a management and technical perspective respectively.
41. **WP5.2 Dissemination**: General external communications will be through regular posts on a dedicated blog for the project, linked to the C4DM and Sound Software sites and Twitter/RSS feeds. Project deliverables will be published on the project site, archived by Sound Software, and shared with the DCC and with other interested parties via JISC channels (e.g. Strand and Programme Meetings, Research Data Management Forum events, direct engagement with other funded projects). Outputs from the project will also be disseminated at conferences and workshops such as the annual Digital Music Research Network meeting, hosted at Queen Mary University of London in December, and the International Digital Curation Conference.
42. **Deliverable D5.1 (M1)**: Project site and feed
43. **Deliverable D5.2 (M6)**: Blog posts, reports, budgets, plans, etc., as required by JISC

3 Engagement with the Community

44. The project is to be carried out directly embedded within the research group at C4DM. The PI (Dixon) and first CI (Plumbley) are very experienced members of the research community in this field and the second CI (Cannam) is experienced in developing and managing the software lifecycle in the research environment. We aim to select for the RA role from individuals with real research experience at the C4DM or another similar centre. The number of everyday interactions with active researchers in this field will be very high, and we are confident of ensuring that the needs of the immediate research community will be taken into account at every stage. More concretely, WP1.2 and WP1.3 call for direct engagement with researchers currently and recently active in C4DM, and in WP2.4 the system implemented will be tested with a pool of active researchers, with feedback used to inform the technical recommendations for WP2.5.
45. The outcomes of this project will be taken on board by the Sound Software project (funded until 2014). This will continue to work with JISC and the DCC to assist in formulating good practice and providing resources to other research groups in this field.
46. Communication with community members outside the C4DM during the lifecycle of this project is to be carried out via regular posts on a dedicated blog, and publication of all deliverables through the same site. This will be linked to the C4DM and Sound Software sites and Twitter/RSS feeds. Other community engagement will take place through JISC Programme and Strand Meetings, DCC Research Data Management Forum events, and direct engagement with other funded projects in this Strand. We will also give a presentation on the project at the Digital Music Research Network (DMRN) one-day workshop in December 2011, to introduce the project and raise awareness of research data management issues. Other possible dissemination venues are the UK e-Science All Hands Meeting, the International Digital Curation Conference, and TransferSummit. Announcements of major project outcomes will be distributed to the DMRN mailing list.