Managing Research Data and Software: A Digital Music Research Viewpoint

Mark D Plumbley
Centre for Digital Music
Queen Mary, University of London
Overview

• Introduction & Motivation

• Our Projects
  • Centre for Digital Music
  • Queen Mary University of London

• Conclusions
Dream: “Ideal” Research Pipeline

Researcher A (“Producer”)
• Read background papers
• Do own research
• Publish paper X

Researcher B (“Consumer-Producer”)
• Read paper X
• Understand/reproduce results in paper X
• Do more research building on X
• Publish paper Y that cites X / produce product that uses X
... and so on.
Real Research Pipeline

Researcher A ("Producer")
• Read background papers
• Do own research (including lots of coding)
• Publish paper X (not enough space for the data and/or code)

Researcher B ("Consumer-Producer")
• Read paper X
• Can’t reproduce or use results in paper X
• Tear out hair
• Give up / do something else

NB: A and B may be in same group (or same person later!)
Reproducible Research

Idea: researchers should be able to reproduce the work of others.

Research used to be “reproducible” from the paper alone. In audio & music research, methods are now too complex. The paper is not enough: need data, algorithm, parameters, ...

So, we need

• The paper (ideally Open Access)
• The data (ideally Open Data)
• The software (ideally Open Source)

Well-known example: WaveLab (Buckheit & Donoho, 1995)

But in audio & music research, few people do this. Why?
Research software in practice

We carried out a Survey of UK audio and music researchers*. 82% developed software, but only 39% of those took steps to reproducibility, and only 35% of those published any code. Only 11% tried to be reproducible and published the code.

(Also: 51% said their code never left their own computer)

* - Oct 2010-Apr 2011, 54 complete + 23 partial responses. For these figures we considered 72 responses.
Why don't we publish code & data?

Our survey suggested:

- Lack of time
- Copyright restrictions
- Potential for future commercial use

Other factors (UK Research Information Network, 2010):

- Lack of evidence of benefits
- Culture of independence or competition
- Quality concerns (self-taught programmers)

Also: it takes effort early in the research cycle; hard to find time/motivation after the paper is published.
Reasons we don't like to admit?


Does this cut both ways?

Can we improve quality by helping people prepare to share?

Barriers to publication and reuse

- Lack of education and confidence with code
- Lack of facilities and tools
- Lack of incentive for publication
- Platform incompatibilities

These are barriers to publication of code.

Related issues for data.
Centre for Digital Music

- World-leading research into digital technologies for new understanding and innovation in music and audio.
- ~60 people: 11 academics, 35 PhD students, 14 researchers
- Research funding: over £17 M since 2007
- Teaching: BEng Audio Systems Eng, MSc Digital Music Proc
- Regular international visitors
- Software: Sonic Visualiser, SoundBite, ...
- Partners: BBC, last.fm, FXpansion, Yamaha, ...

EPSRC
Engineering and Physical Sciences Research Council

BBC R&D

last.fm
the social music revolution

YAMAHA

soundsoftware.ac.uk
Our projects

Centre for Digital Music (C4DM)

- SoundSoftware.ac.uk - 2010-2014
  Sustainable Software for Audio & Music Research
- Sound Data Management Training (SoDaMaT) Jun 2012 - Jan 2013

Queen Mary, University of London – College Level

- Research Data Curation: Project Board
  Part of QMUL “IT Transformation” project
Funding from EPSRC (2010-2014) to:
• support the development and use of software and data
• to enable high quality research
• in the audio and music research community

How?
• Developers to make research software robust & usable
• Training for researchers in writing their own code
• Promote software development in research projects
Sustainable Management of Digital Music Research Data (SMDMDRD)

- October 2011 - May 2012
- Pilot project: set up a research-group research data repository
- Chose DSpace for repository:
  - Easy to install
  - Standards compliant
- Tried U. of Oxford's DataStage to link to DSpace, but not live
- Command-line tool created to upload data to the repository using SWORDv2 protocol

C4dm.eecs.qmul.ac.uk/rdr/xmlui/
Sound Data Management Training (SoDaMaT)

- June 2012 - January 2013
- Project to create discipline-specific RDM training materials for C4DM
- Materials to be targeted at postgraduates and researchers
- Tutorials presented at digital audio conferences (ISMIR 2012 and DAFx 2012)
- Training materials to be published on Jorum
- Online training materials
- https://code.soundsoftware.ac.uk/projects/sodamat
SoDaMaT: Example RDM failures

Subject: Recovery of Overwritten Hard Disk Data

Hi, a friend of mine just overwrote two months of her PhD thesis with an older version. I know recovery of overwritten data is possible, but wonder if I'd need special hardware to do it. Does anyone know something about this?

Thank You.

5 October 2005 Linux Forums - http://tinyurl.com/8t7uaop
QMUL Research Data Curation: Project Board

- Jan – Nov 2013
- Timely – fitted EPSRC requirement for RDM policy
- Included input from academics and library
- PRINCE2 project (I’m a “Senior User”)

Executive
Evelyn Welch  Vice-Principal for Research and International Affairs

Senior User
Gerry Leonard  Head of Research Resources
Sarah Molloy  Research Support Librarian
Paul Smallcombe  Records & Information Compliance Manager
Aine McKnight  SMD (Blizard)
Mark Plumbley  S&E (Electronic Engineering and Computer Science)
Martin Dove  S&E (Physics and Astronomy)
Isabel Rivers  HSS (English and Drama)
David Van Heel  SMD (Blizard)
Michael Barnes  SMD (William Harvey)

Senior Supplier
Chris Day/Research AD  IT Services

Project Manager
Paul O’Shaughnessy/ TBA  IT Services
Timeline

- Jan: DCC visit
- Feb: PM in place
- Mar: Project Start
- Apr: Objectives agreed
- DCC Cardio
- Draft RDM Policy
- Data Management Roadmap
- Policy awareness and promotion
- Project objective #1
- Approved RDM policy
- Data repository requirements
- Policy objective #2
- QM DMP online
- Data repository requirements
- Project objective #3
- Web pages
- Costing Tool
- Project objective #4
- Project End
- Project objective #5
- Project End
Grant costing template

INSTRUCTIONS

1. Use the buttons above to navigate through the calculator
2. First complete the questionnaire - you need to fill out all the yellow shaded cells
3. Based on your answers to the questionnaire, a cost justification will be created. Click on the navigation button to see the completed justification
4. Check the wording to make sure your answer and calculations have come across correctly
5. Select the cell containing the justification, Right click your mouse and select copy
6. Open a blank document in Word
7. Select paste special 'Text Only'
8. Select the IT Cost Summary tab
9. Check the numbers to make sure your answer and calculations have come across correctly
10. Select the table cells, Right click your mouse then select copy
11. Go back to your document in Word
12. Select paste
# Delivery Plan

## QMUL Research Data Management

### Roadmap Delivery Plan

<table>
<thead>
<tr>
<th>RDM Theme</th>
<th>Phase 1: Enabling RDM</th>
<th>Phase 2: RDM Skills &amp; Tools</th>
<th>Phase 3: Best Practise</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Policy Review</strong></td>
<td>Policy alignment</td>
<td>Set up Policy Review</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Definition of Data, Data types, Roles &amp; Responsibilities</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Business Planning</strong></td>
<td>Inclusion of research data in Publication process</td>
<td>Demand Forecasting</td>
<td>Integrate to REF reporting</td>
</tr>
<tr>
<td></td>
<td>Allocations of resource to RDM training</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Registration of Research</strong></td>
<td>Tools for Registration of Research</td>
<td>Integration to Grant applications, publications and resource management systems</td>
<td>Encourage registration of unfunded research</td>
</tr>
<tr>
<td><strong>Consultancy Support</strong></td>
<td>RDM Services and Consultancy</td>
<td>Tools for curation, preservation, metadata &amp; obsolescence</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Catalogue of IT Services for Researchers</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tools to support Data Management Planning</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>IT Planning</strong></td>
<td>Align Skills to RDM Requirements</td>
<td>Forward planning and managing technology change</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Plan for appropriate technologies</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Costs &amp; Sustainment</strong></td>
<td>Build simple transparent cost model grant application</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Build consistent, scalable, mechanisms for re-charge</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Repository</strong></td>
<td>Build Research Data Repository</td>
<td>Build links to external Data Repositories standards &amp; tools</td>
<td>Develop statistics and analytical for REF</td>
</tr>
<tr>
<td><strong>RDM Support</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Develop central infrastructure, networks, storage and processing capacity, back-up and continuity facilities, security and integrity</td>
<td>Build sharing capabilities, collaborative tools and specialised services</td>
<td></td>
</tr>
<tr>
<td><strong>RDM Tools</strong></td>
<td>Develop guidance for RDM utilising independent external tools</td>
<td>Tools for curation, preservation, metadata &amp; obsolescence</td>
<td>Continual development of tools and best practise</td>
</tr>
</tbody>
</table>

[(soundsoftware.ac.uk)](soundsoftware.ac.uk)
Some remaining issues

• What is data?
• One idea: “Anything that you need to validate the research in a published paper, that isn’t in the paper itself”
• So it could be:
  • Survey result (did I ask the participants if I could share?)
  • Music tracks (who owns those?)
  • Software (but my university owns the IP?)
  • The Internet (hmm ...?)
• So, consider what you mean by “validate” above
• Also: Data and Software people don’t always talk – why?
Putting it all together

What we’re trying to do:

• Create an Open Access, Reproducible Research culture
• Get help from the library – provides central service
• Get researchers to think about Data, Software and Reproducible Research right from the start
• Training in research software dev. and data management
• Collaborative environment to develop & share code
  Write code expecting other people will read it
• Refer to data somebody else owns
• Reproducible Research Repository: link paper-software-data
Conclusions

• Data and software is important for our research
• Impossible to validate our research without it
• Researchers need help to develop software and manage data
• C4DM: RDM server, training
• QMUL: RDM project (helped by EPSRC)

• Make research work better!