soundsoftware.ac.uk

Managing Research Data and Software: A Digital Music Research Viewpoint

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

EPSRC
Engineering and Physical Sciences Research Council

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

(Buckheit & Donoho, 1995; Vandewalle et al, 2009)

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?

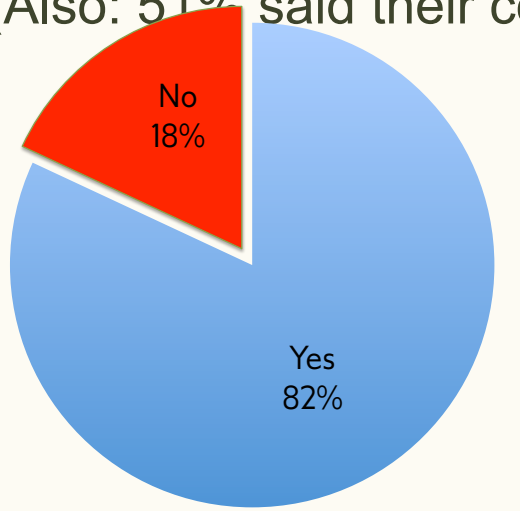
Research software in practice

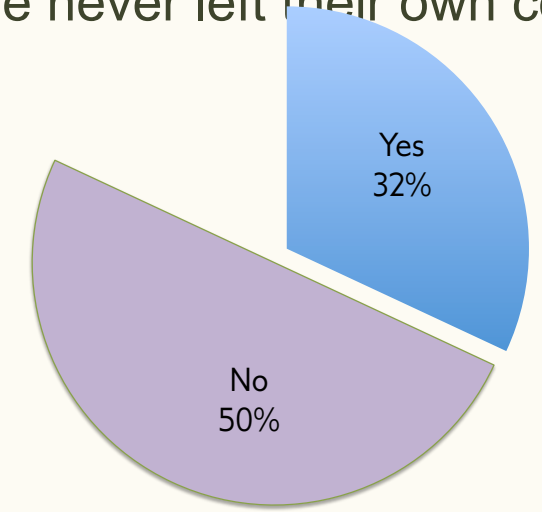
We carried out a **Survey of UK audio and music researchers***.

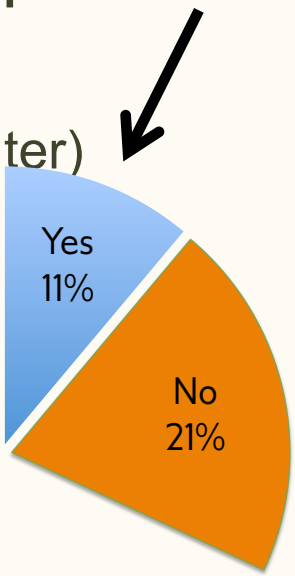
82% developed software, but only 39% of those took steps to reproducibility, and only 35% of *those* published any code

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?

J M Wicherts, M Bakker and D Molenaar, 2011, *Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results*, PLoS ONE

Does this cut both ways?

Can we improve quality by helping people prepare to share?

Barriers to publication and reuse

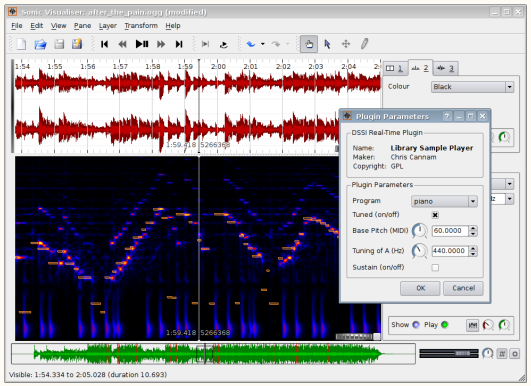
- Lack of education and confidence with code
- Lack of facilities and tools
- Lack of incentive for publication
- Platform incompatibilities

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
- Conferences: ISMIR 2005, ICA 2007, MPEG 2009, CMMR 2012
- Teaching: BEng Audio Systems Eng, MSc Digital Music Proc
- Regular international visitors
- Software: Sonic Visualiser, SoundBite, ...
- Partners: BBC, last.fm, FXpansion, Yamaha, ...



BBC R&D

last·fm the social music revolution

YAMAHA

EPSRC Engineering and Physical Sciences Research Council

soundsoftware.ac.uk

Our projects

Centre for Digital Music (C4DM)

- SoundSoftware.ac.uk - 2010-2014
  Sustainable Software for Audio & Music Research
- Sustainable Management of Digital Music Research Data (SMDMRD) - Oct 2011 - May 2012
- 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

SoundSoftware.ac.uk

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 (SMDMRD)

- 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

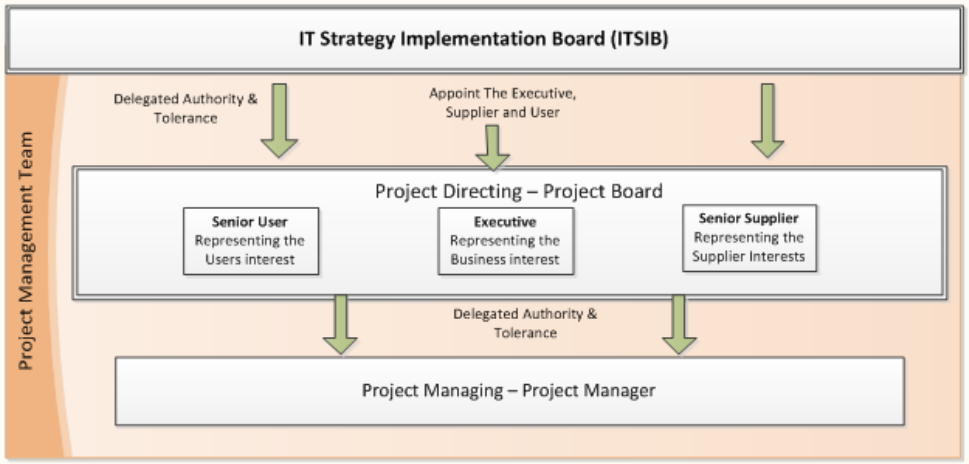
Subject: Recovery of Overwritten Hard Disk Data

> 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. Dos anyone know something about this ?
>
> Thank You.

5 October 2005 Linux Forums - http://tinyurl.com/8t7uaop

QMUL Research Data Curation: Project Board

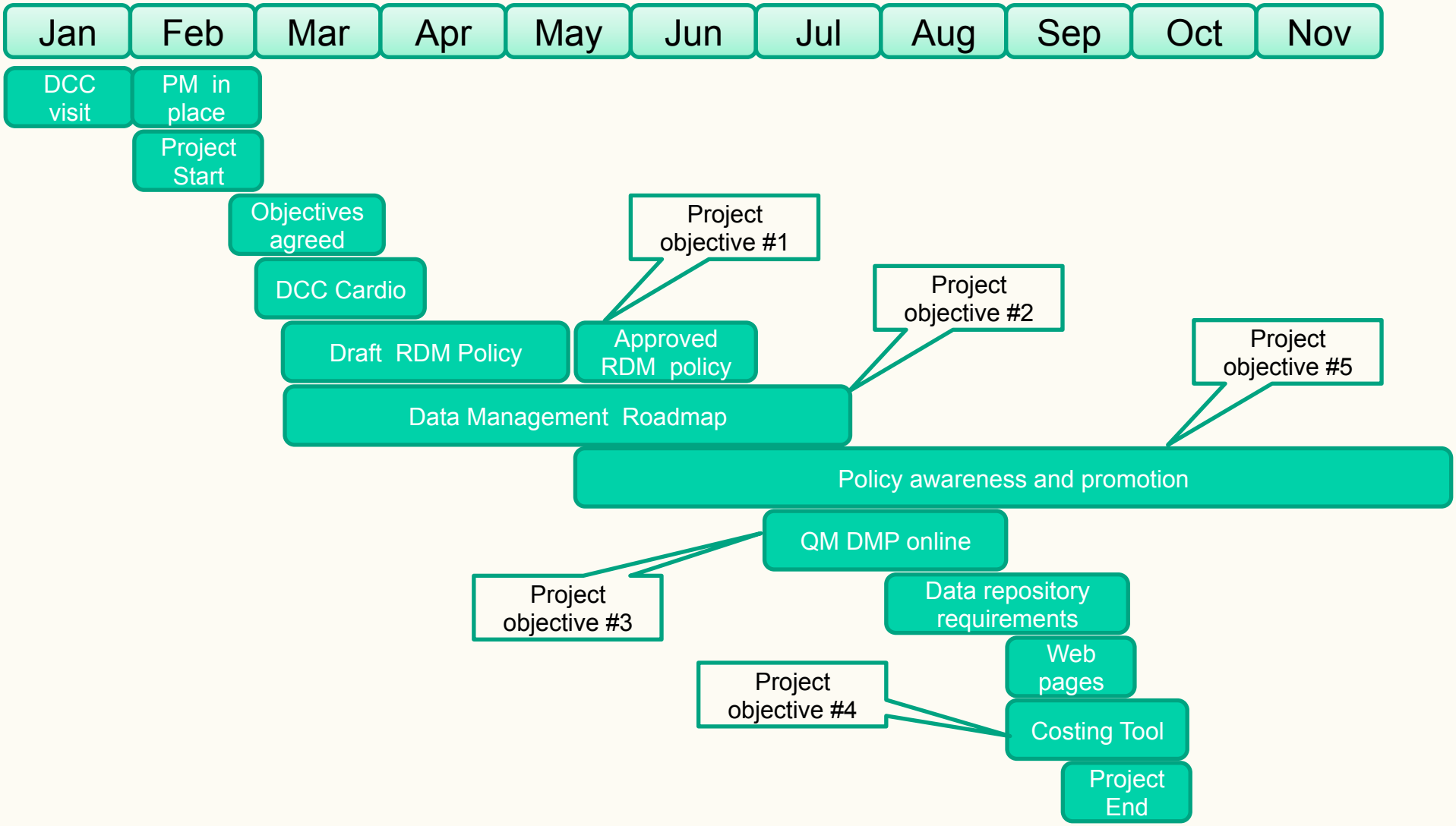
- Jan – Nov 2013
- Timely – fitted EPSRC requirement for RDM policy
- Included input from academics and library
- PRINCE2 project (I 'm a "Senior User")



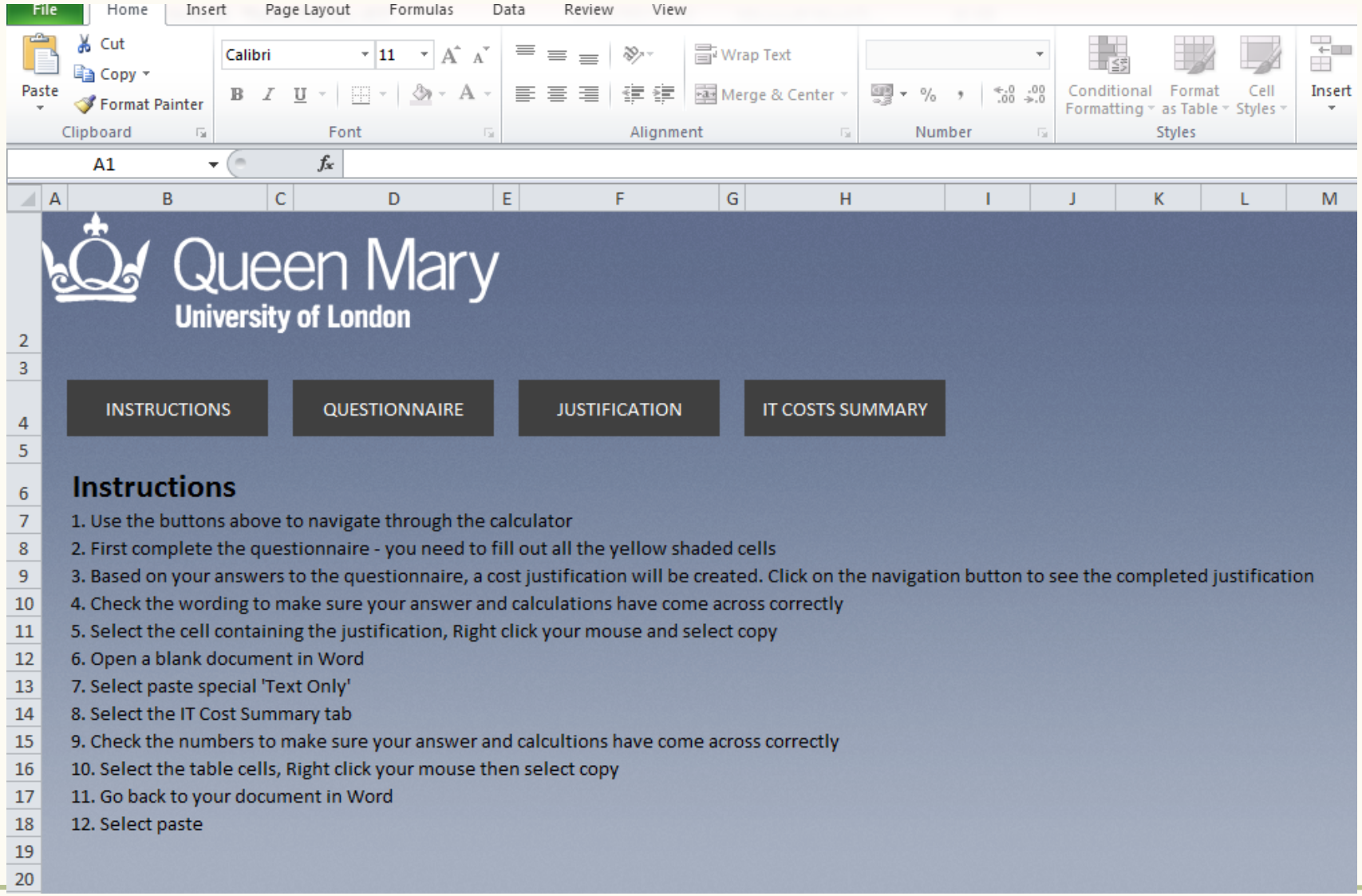
Executive	Evelyn Welch	Vice-Principal for Research and International Affairs
Senior User	Gerry Leonard	Head of Research Resources
	Sarah Molloy	Research Support Librarian
	Paul Smallcombe	Records & Information Compliance Manager
	Aine McKnight	SMD (Blizard)
	Mark Plumbley	S&E (Electronic Engineering 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
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
DCC visit
PM in place
Project Start
Objectives agreed
DCC Cardio
Draft RDM Policy
Approved RDM policy
Project objective #1
Project objective #2
Project objective #5
Data Management Roadmap
Policy awareness and promotion
QM DMP online
Project objective #3
Data repository requirements
Web pages
Project objective #4
Costing Tool
Project End

Grant costing template



Delivery Plan

QMUL Research Data Management
Roadmap Delivery Plan

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

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!