soundsoftware.ac.uk

Managing Research Data

Steve Welburn
Centre for Digital Music
Queen Mary, University of London

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

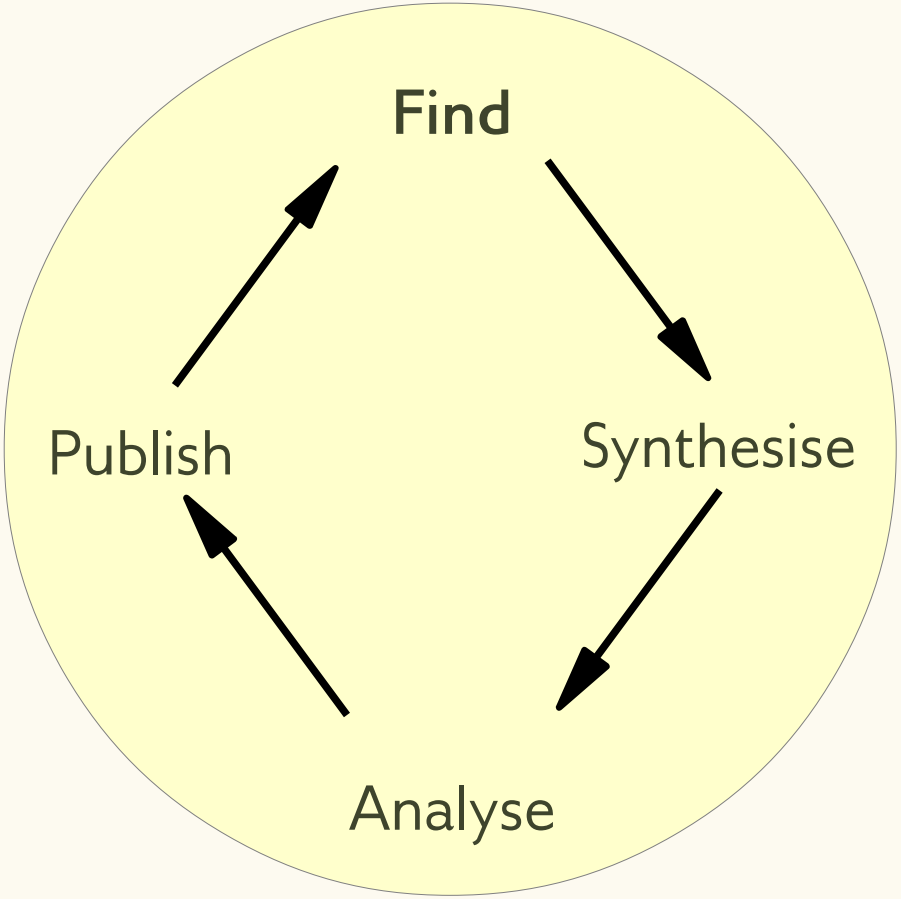
Overview

- Active Data Management
  - How can you lose data
  - How can you avoid it!
  - Organizing data
  - File Formats
- After Your Research
  - Archiving and publishing data
  - Ownership
  - Licenses

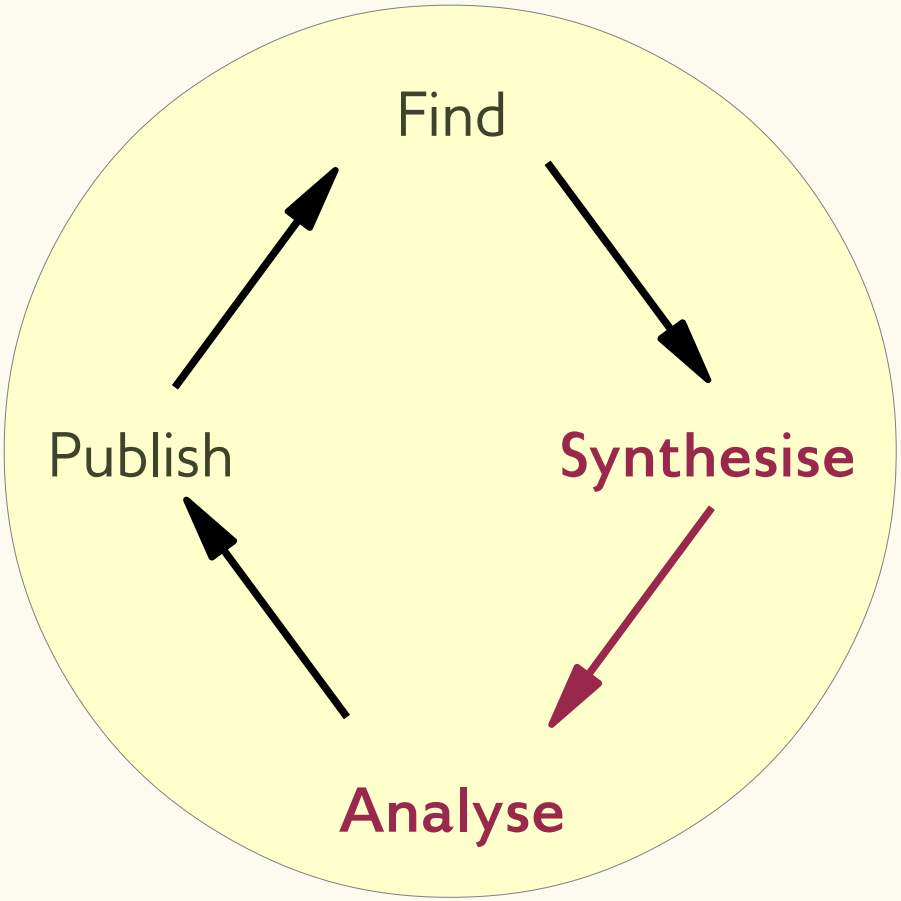
Benefits

- Meeting funder requirements (e.g. EPSRC)
- Reduce risk of losing work
- Possibility of citations based on data
- Opportunities for follow-on research and collaboration
- Good practice!

Research Lifecycle



Active Data Management



What is Research Data ?

- Digital data you use in your research
  - Reference/Standard datasets
  - Interview transcripts
  - Annotations for audio / video files
  - Circuit diagrams
  - Downloaded papers
  - Software you've written (Version Control!)
  - Citation management
  - etc...

Data Analysis

- If an analysis is particularly involved, it should be designed as a series of analysis steps
- This allows you to rerun only the affected bits of the analysis after changes
- Once it works, you can still use a script to run all steps of the analysis if it's designed as separate steps
- You will (probably) need to save intermediate results to run in multiple steps
- That's more data to manage
- Record the version of the relevant code with the intermediate results

Data Analysis

- Running small incremental analyses also makes it easier to spot where bugs occur
  - Avoids discovering bugs that were early in the analysis only after you've completely processed 1000 files.
  - If you already know steps 1-3 work and step 4 shows errors, step 4 needs fixing
  - If you just ran all of 1-4 then you have a bigger job finding the problem
  - The earlier you find bugs, the less time will be wasted!

Managing Data

- Source code and scripts
  - Use version control
- Citations
  - Desktop applications
    - e.g. Reference Manager, EndNote
    - Site licence for EndNote on QMUL machines
  - Online reference management
    - e.g. Zotero, Mendeley
- Everthing else
  - ...is what we're going to concentrate on

Organising Folders

- Folders should contain either code or data - not both
  - Projects
    - Project1
      - Experiment 1
        - Code
        - Data
      - Experiment 2
      - Experiment 3

Organising Folders

- Projects
  - Project1
    - Data
    - Experiment 1
    - Experiment 2
    - Experiment 3

File Formats

- Open file formats vs. proprietary formats
  - Open formats allow content to be recovered based on the documentation for the format
  - Proprietary formats may have no publicly available documentation – if you can't find software that reads the format the data will be inaccessible
- Lossy vs. lossless formats
  - There is no guarantee that future decoders will decode lossy data in exactly the same manner as current decoders
  - Document how you decode data

Naming Files

- File Names
  - Should be meaningful and brief
  - Should not depend on the folder structure
    - Files may be copied to different folders
  - Can indicate provenance
    - Who ? When ? Why ? How ?
  - Date created / modified from the OS is unreliable information as it may change when files are copied

Example:

- Bad: piano.wav
- Good: sjw_e12_20120829_piano.wav

Document

- If data allows is accessed at a later date, will it be usable ?
- Will people understand:
  - Why you created it ?
  - What the data is useful for ?
  - What column 27 in table 15 actually means ?
  - How the data was created (e.g. which algorithm) ?
  - What the source data was on which this data is based ?

- If you return to your data to check something at the end of your research, will **you** understand the data ?

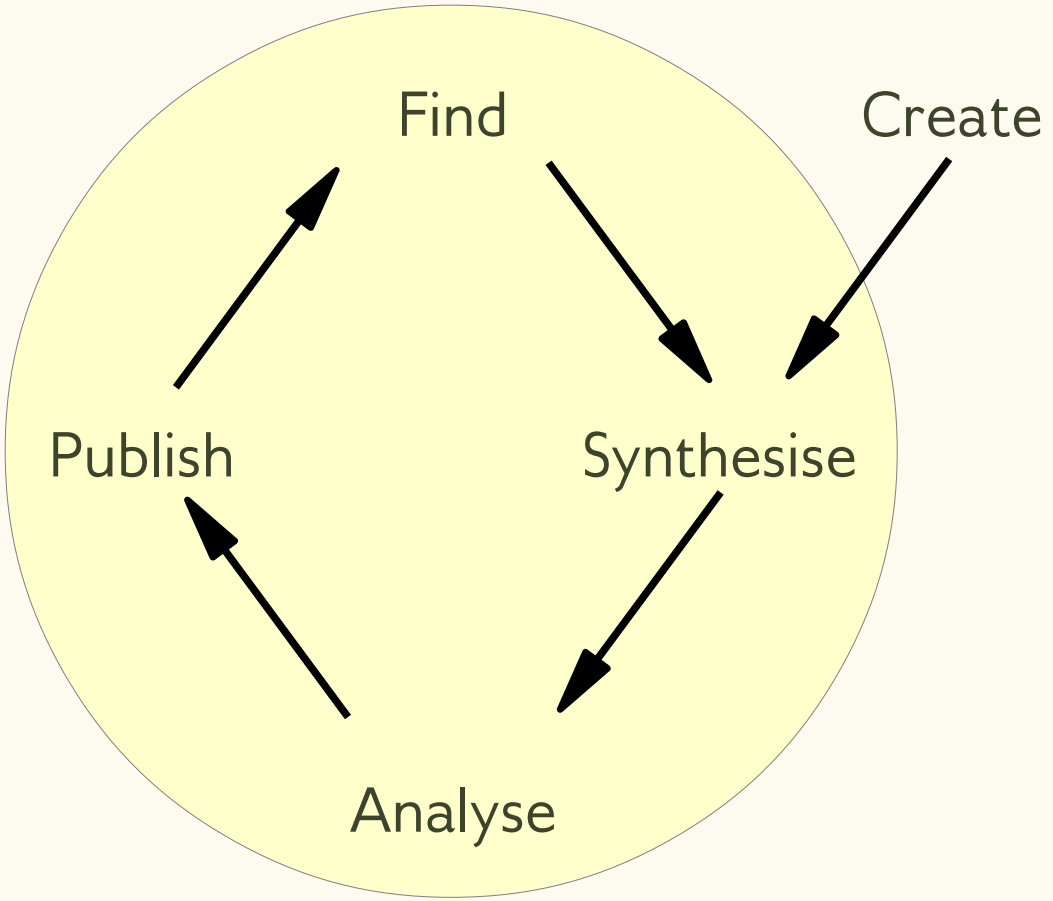
Documenting Data

- Metadata (data about data) should be provided to describe:
  - Contents – what is the data ?
  - Purpose – why is it useful ?
  - Provenance – how was the data created ?
  - Licence – how can it be used ?
  - Audience – who might be interested ?
- Metadata does not need to be structured, a README file explaining the file contents is sufficient.
- Keeping documentation with the data means it is there if you give someone a copy of the data

Document Parameters Used

- If you process data, then document all parameters that are used.
- If you use default parameters, document the values that those parameters take – other versions of the processing may provide different default values.
- Using scripts to process data can include the parameter values in the scripts
  - And putting the scripts in version control records those values

Research Lifecycle



Backing Up Data

Stolen laptop had PhD research

Thirty-five minutes spent in Langley's Willowbrook Shopping Centre cost a Surrey woman much more than she had anticipated.

Langley RCMP say that while she was shopping from 1-1:35 p.m. last Monday, someone broke into her vehicle and stole a number of items, including a Mac iBook laptop containing the research she had compiled as she worked towards her PhD.

"All that information was on that computer and she has no back-up file," said Langley RCMP spokesman Cpl. Brenda Marshall.

19 March 2008 Surrey Leader - http://tinyurl.com/9hmtlv4

Just a working copy

- Risks include:
  - Overwriting your data
    - Manually
    - By running a buggy algorithm
  - Deleting folders to salvage disk space
  - Deleting the wrong file
  - Virus attack
  - Letting other people use your computer

- Keep backups!

Happiness is the return of a stolen computer, with data intact

Never has a man been so happy to see a computer full of data spreadsheets.

Claudio De Sassi's world fell apart when a car containing almost three years work towards his PhD was stolen two weeks ago.

De Sassi, a Canterbury University academic, could not hide his joy yesterday as police reunited him with his stolen laptop and backpack.

27 May 2010 The Press, NZ - http://tinyurl.com/38sznnh

The Lost Laptop Problem

- 2010 Ponemon Institute report for Intel re. US laptops
  - On average, 2.3% of laptops assigned to employees are lost each year
  - In education & research that rises to 3.7%, with 10.8% of laptops being lost before the end of their useful life (~3 years i.e. within 1 PhD of allocation!
  - 75% lost outside the workplace

- Very similar results from 2011 European report!

Intel 2010 - http://tinyurl.com/8c9m4bn

Laptop Reliability

- 2011 PC World Laptop Reliability Survey from 63,000 readers:
  - 22.6% had significant problems during the product's lifetime
  - Of which...
    - 19% had OS problems ~1 in 25 of all laptops
    - 18% had HDD problems ~1 in 25 of all laptops
    - 10% PSU problems ~1 in 50 of all laptops

PC World 2011 - http://tinyurl.com/876qza5

Hard Disk Failures

- *Failure Trends In A Large Disk Drive Population*
  - Usenix conference on File and Storage Technologies 2007 (FAST '07)
  - Eduardo Pinheiro & Wolf-Dietrich Weber, Google Inc.

- Data collected from over 100,000 disk drives at Google
- As part of repairs procedures:
  - ~13% of disk drives replaced over 3 years
  - ~20% of disk drives replaced over 4 years

Article: http://tinyurl.com/octz6b

Backups on Removable Media

- Data on laptop and backups on removable media

- Risks:
  - Losing or misplacing the media
  - Forgetting to label DVDs
  - Keeping the backup with the laptop

- Mitigation:
  - Catalogue your backups
  - Store your backups apart from your computer

Thugs steal Christmas, doctoral dreams

A tiny television sits where a big screen used to, and a Christmas tree stands with little underneath it...

Even worse than the gifts, the crooks stole a MacBook Pro laptop and a LaCie hard drive.

The hard drive had ... her dissertation and nearly seven years of research for her doctoral degree she was set to finish in a few weeks.

Osuna had everything backed up on a separate hard drive in a safe, but burglars made off with that too.

"All I could think about is that all that time is gone, all that effort, everything is gone," Osuna said.

22 December 2010 KRQE - http://tinyurl.com/9a5j56f

Where To Keep Your Data

- Keeping copies of data in separate locations helps you avoid losing your data.
- A separate location could be:
  - removable media (e.g. USB stick, DVD-R)
  - a network drive
  - in “the cloud”
- Although it's easy to do backups on physical media, network backups usually provide a better service.
- Remember that if you delete the local copy because you have a backup you are back to only one copy existing!

Where To Keep Your Data

- Commercial remote storage solutions (e.g. DropBox)
  - Check the T&Cs / SLA
  - Cost money
  - Not openly accessible on the web
  - No control over how data is stored
  - No control over physical location of data
  - Risk of lock-in
  - Bandwidth restrictions

JISC/DCC Curation In The Cloud : http://tinyurl.com/8nogtmv

Cloud Storage - Google

When you upload or otherwise submit content to our Services, you give Google (and those we work with) a worldwide license to use, host, store, reproduce, modify, create derivative works (such as those resulting from translations, adaptations or other changes we make so that your content works better with our Services), communicate, publish, publicly perform, publicly display and distribute such content. The rights you grant in this license are for the limited purpose of operating, promoting, and improving our Services, and to develop new ones. This license continues even if you stop using our Services (for example, for a business listing you have added to Google Maps).

1 March 2012 Google Terms of Service : http://tinyurl.com/89dc9fa

Cloud Storage - Microsoft

When you upload your content to the services, you agree that it may be used, modified, adapted, saved, reproduced, distributed, and displayed to the extent necessary to protect you and to provide, protect and improve Microsoft products and services. For example, we may occasionally use automated means to isolate information from email, chats, or photos in order to help detect and protect against spam and malware, or to improve the services with new features that makes them easier to use. When processing your content, Microsoft takes steps to help preserve your privacy.

19 October 19 2012 Microsoft services agreement : http://tinyurl.com/8e4kucy

Where To Keep Your Data

- Institutional Network Storage
  - May be available already
  - Should intend to support your research
  - May be difficult to find out about!

Laptop Stolen From OSU Doctoral Student

...her car was broken into and her chrome Mac book pro was stolen.

She has a back-up for all but the last six months of research, but the most important part of the research had happened recently.

NBC4i January 06 2011 - http://tinyurl.com/bmybv9x

Schedule Backups

- Backups are no use if they are out of date
- Get into the habit of backing up your data regularly
  - How regularly is your choice
  - How much work are you willing to risk losing ?

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.

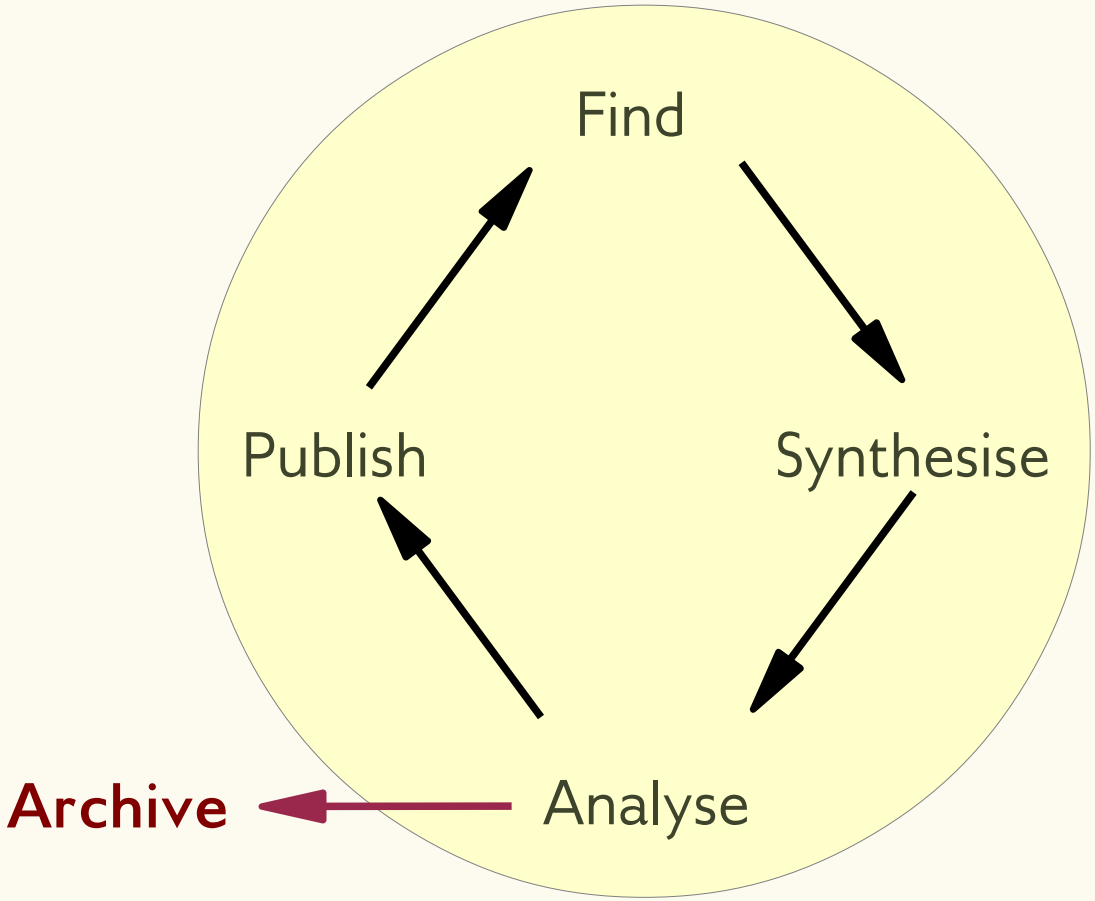
5 October 2005 Linux Forums - http://tinyurl.com/8t7uaop

Remember, to take care if you need to recover from a backup! Even better backup your current state before any recovery takes place!

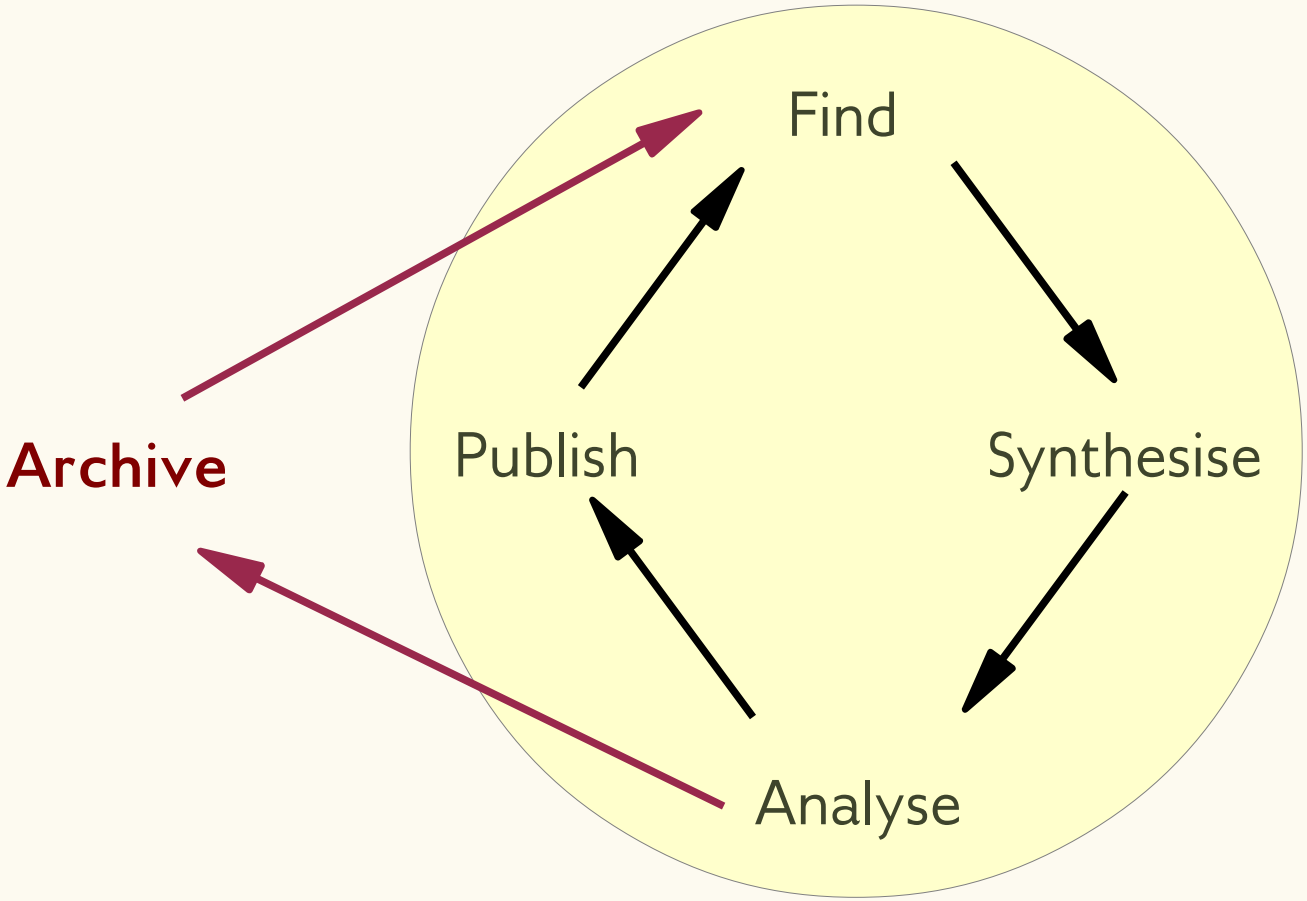
After your research

- At the end of your research you should archive your data for long-term access:
  - for follow-on research
  - to allow validation of your results

Archiving data



Archiving data



Archiving Data

- BBC Domesday Project (1986)
  - Project to do a modern-day Domesday book
  - Used “BBC Master” computers with data on laserdisc
  - Collected 147,819 pages of text and 23,225 photos
  - Media expiring and obsolete technology put the data at risk!

Archiving Data

- Domesday Reloaded (2011)
  - Required emulation of software
  - Images restored from original masters
  - http://www.bbc.co.uk/history/domesday

Lessons We Can Learn...

- To allow long-term access to data
  - Don't use obscure formats!
  - Don't use obscure media!
  - Don't rely on technology being available!
  - Do keep original source material!

Long-term Data Storage

- Disks wear out, and interfaces become obsolete so data should be copied to fresh media at intervals

- Old formats can become unusable
  - Use open formats rather than closed formats
  - Refresh formats to ensure availability

- This is an effort!
  - If possible, let someone else do it by placing your data in an archive which will deal with these issues for you.

Preserve

- Given the number of ways you can lose data, you should take precautions to protect it!

- Will **your** data be available:
  - When you need it ?
  - If someone else needs it ?

Publishing Data

- Publishing data allows other people to
  - Validate your research
  - Check their implementation of your algorithm
  - Produce directly comparable results
  - Combine individual datasets into a good test corpus

- And allows you...
  - To get cited when the data is used

What to publish ?

- Data that will allow others to validate your research
  - Results which are summarised in a publication
  - e.g. the full data behind graphs, tables and statistics

- Data for others to use in their research
  - New datasets which can be used to test new and existing algorithms
  - e.g. annotations for audio datasets and new audio datasets

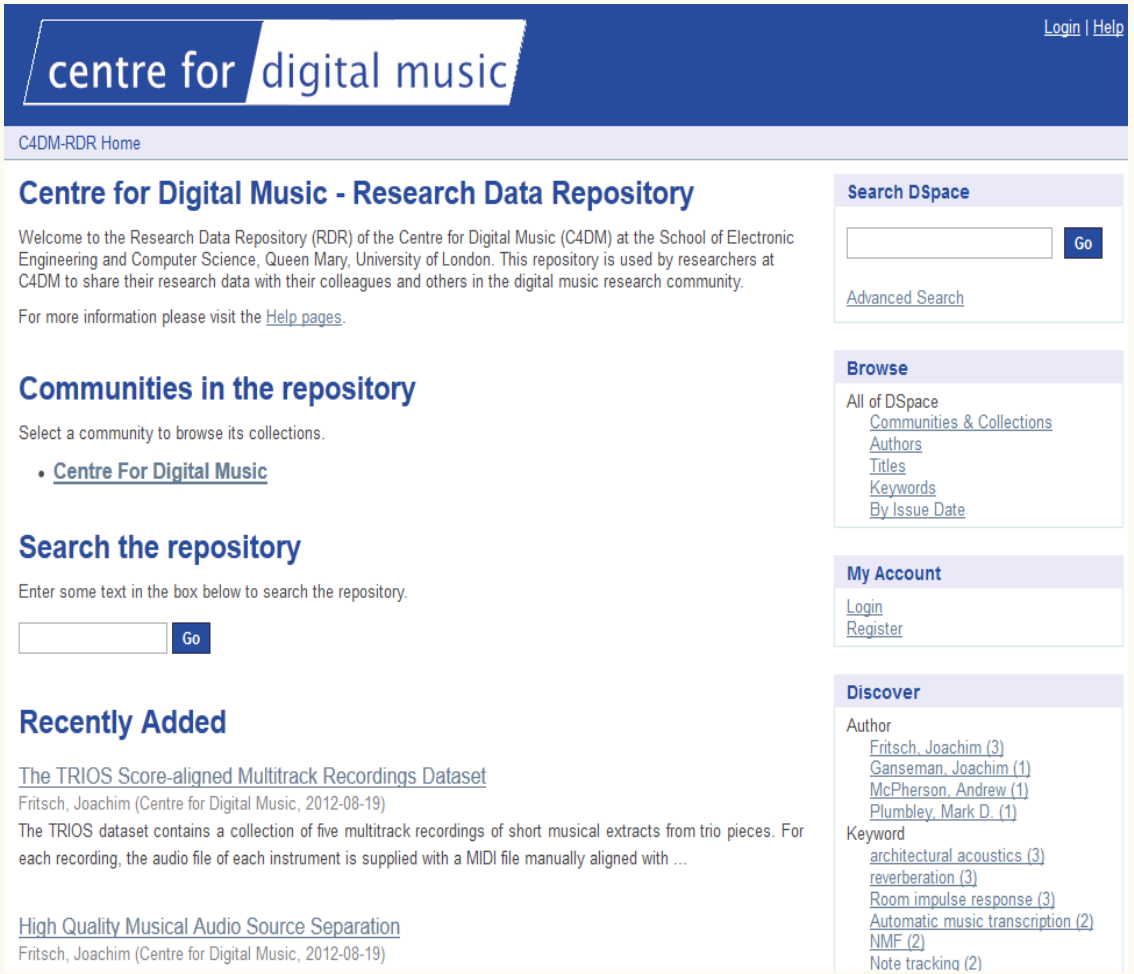
- References to source datasets used in your research
  - e.g. lists of CD catalogue numbers

Where to publish data

- Institutional repository – if one exists!
- Project or research group web-sites
- Journal Supplementary Materials
  - e.g. JASA, JNMR, CMJ
  - Check T&Cs – e.g. JASA ask for copyright to supplementary materials to be transferred to them.
- Web archives – e.g. archive.org for audio files
- Research data sites e.g. figshare.com

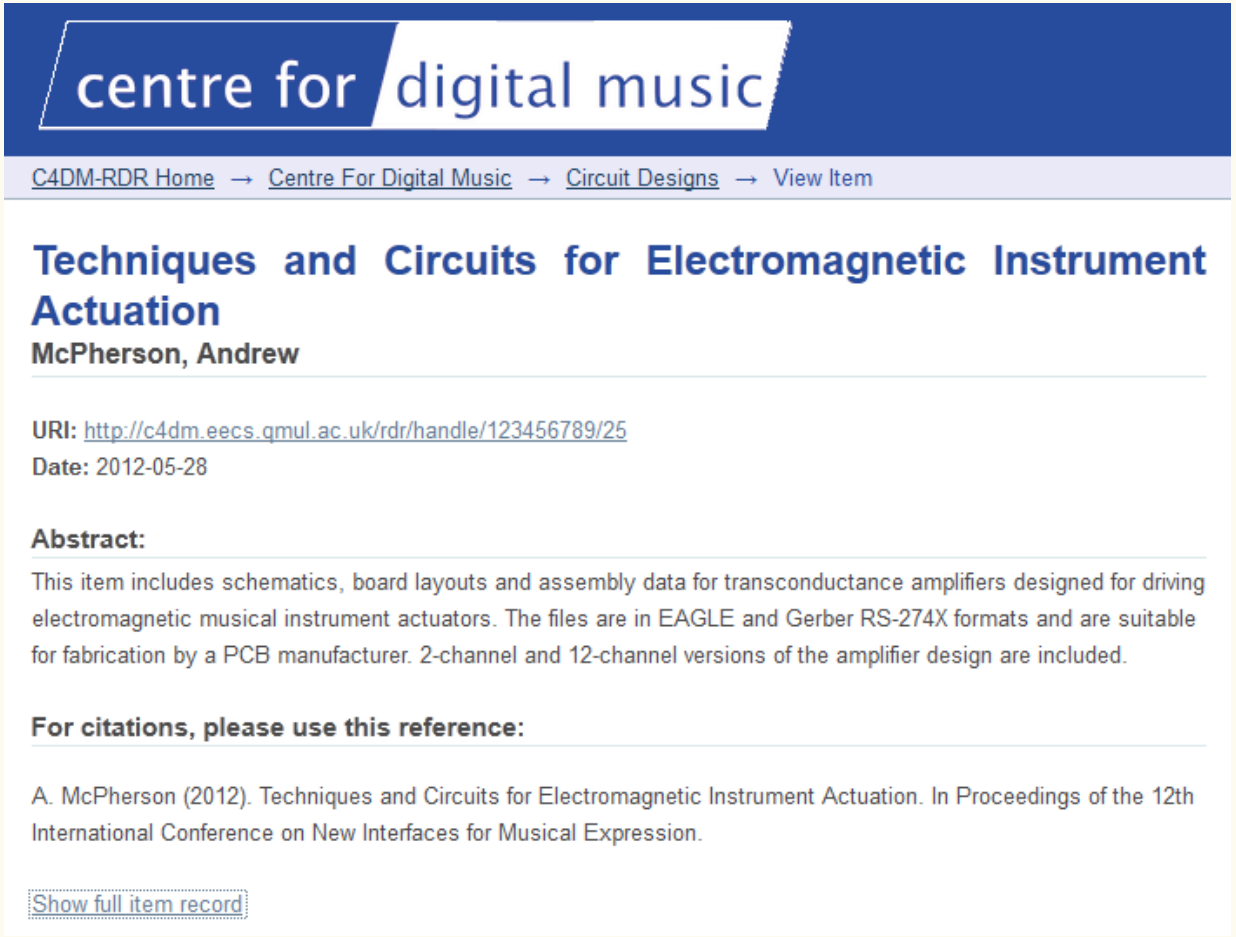
- Talk to a librarian!

C4DM Research Data Repository



http://c4dm.eecs.qmul.ac.uk/rdr

C4DM Research Data Repository


centre for digital music

C4DM-RDR Home → Centre For Digital Music → Circuit Designs → View Item

Techniques and Circuits for Electromagnetic Instrument Actuation

McPherson, Andrew

URI: http://c4dm.eecs.qmul.ac.uk/rdr/handle/123456789/25

Date: 2012-05-28

Abstract:

This item includes schematics, board layouts and assembly data for transconductance amplifiers designed for driving electromagnetic musical instrument actuators. The files are in EAGLE and Gerber RS-274X formats and are suitable for fabrication by a PCB manufacturer. 2-channel and 12-channel versions of the amplifier design are included.

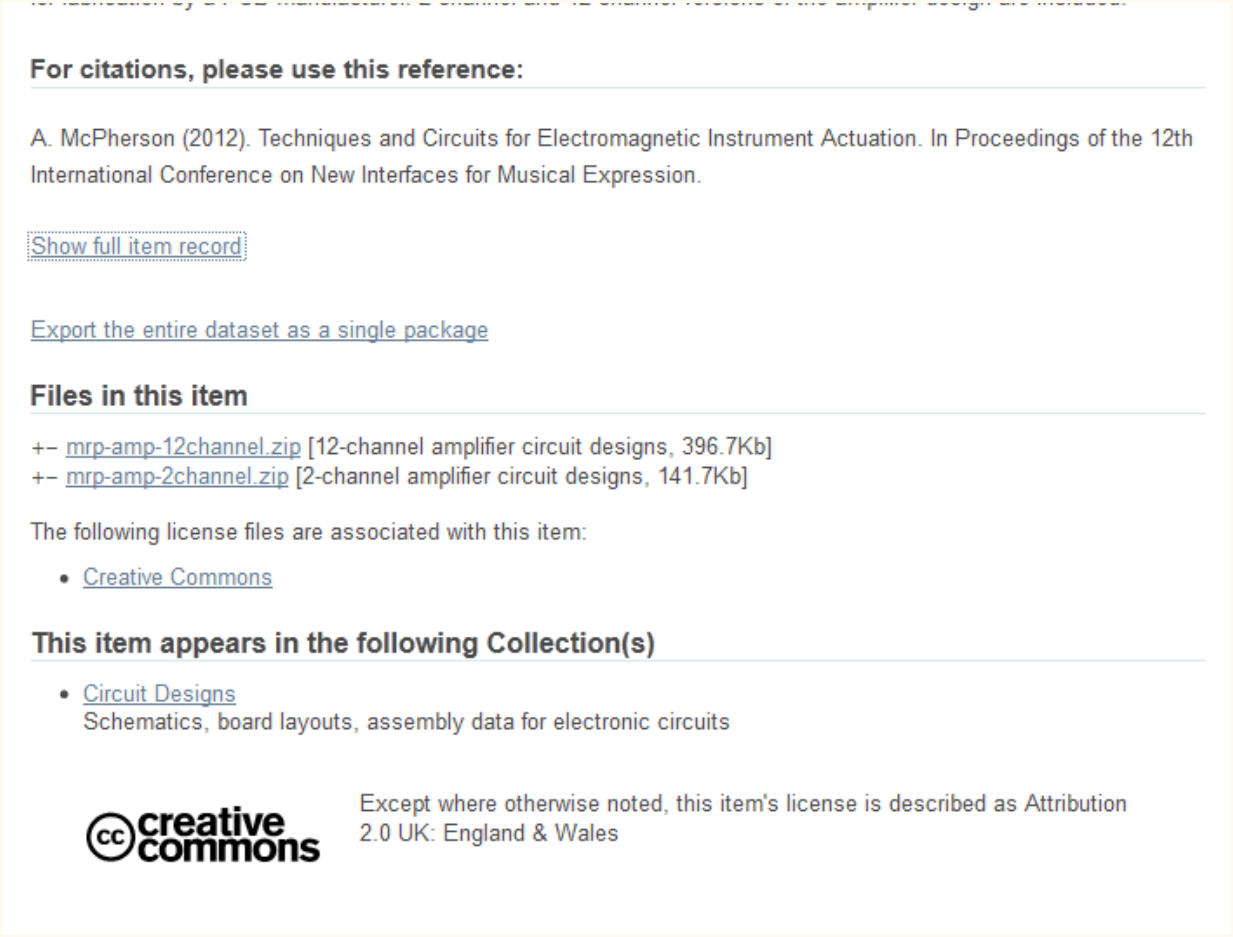
For citations, please use this reference:

A. McPherson (2012). Techniques and Circuits for Electromagnetic Instrument Actuation. In Proceedings of the 12th International Conference on New Interfaces for Musical Expression.

Show full item record

http://c4dm.eecs.qmul.ac.uk/rdr

C4DM Research Data Repository



For citations, please use this reference:

A. McPherson (2012). Techniques and Circuits for Electromagnetic Instrument Actuation. In Proceedings of the 12th International Conference on New Interfaces for Musical Expression.

Show full item record

Export the entire dataset as a single package

Files in this item

+– mrp-amp-12channel.zip [12-channel amplifier circuit designs, 396.7Kb]
+– mrp-amp-2channel.zip [2-channel amplifier circuit designs, 141.7Kb]

The following license files are associated with this item:

- Creative Commons

This item appears in the following Collection(s)

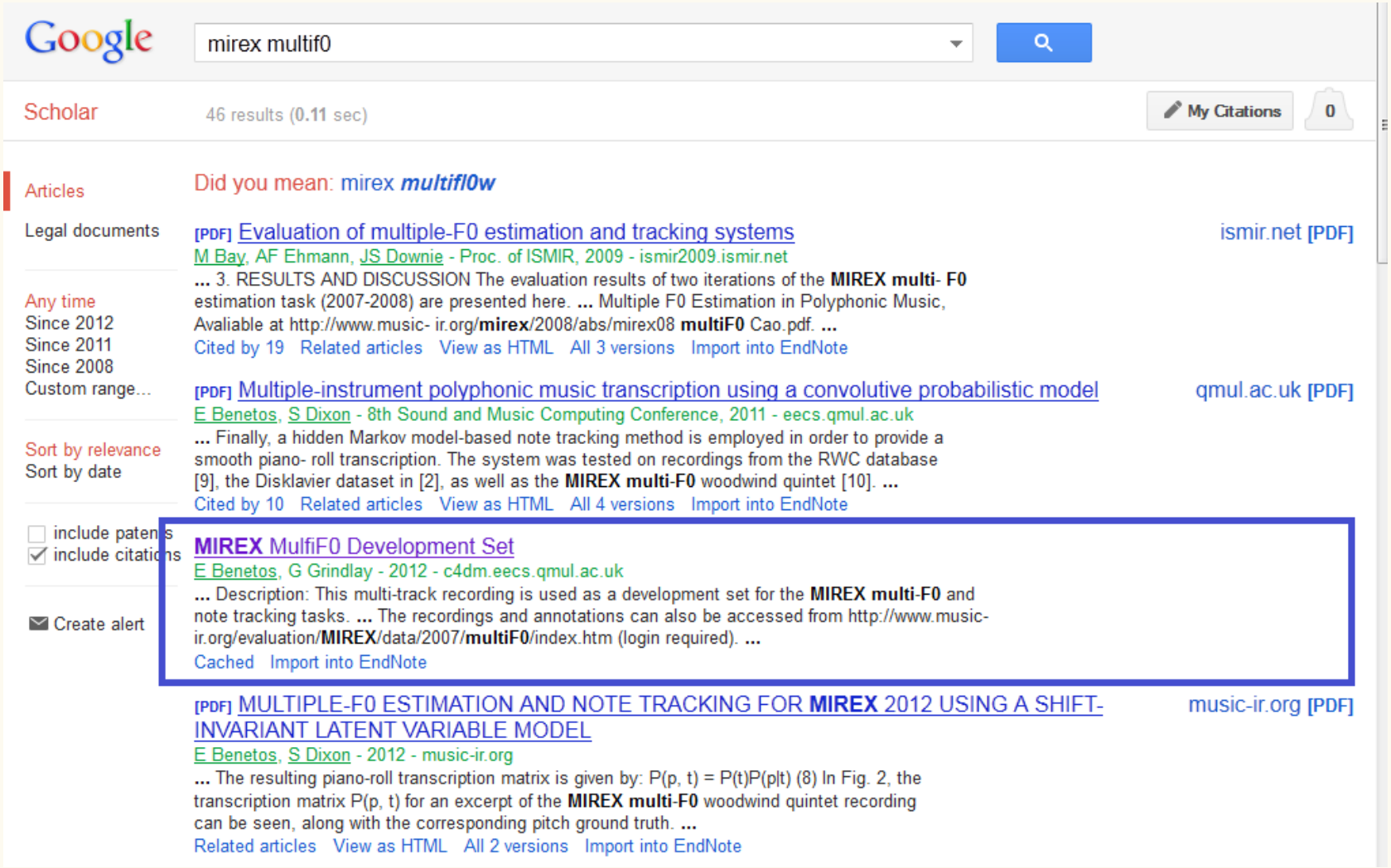
- Circuit Designs
  Schematics, board layouts, assembly data for electronic circuits

creative commons

Except where otherwise noted, this item's license is described as Attribution 2.0 UK: England & Wales

http://c4dm.eecs.qmul.ac.uk/rdr

C4DM Research Data Repository



Licensing Research Data

- If you don't supply a license, you reserve all rights to its use
- Copyright does not exist on factual data itself, only on the "creative" part of the data – e.g. the layout of a spreadsheet

- Some data that is available online has no licence agreement. Which means you actually have no right to use it – email the authors.
- Some data comes with a specific licence – read it and understand it.

- The most common licences for recent data are Creative Commons licences.

Creative Commons Licences

- The standard Creative Commons licences give people the right to use and share your data... with optional restrictions on use:
  - Attribution – it can only be used if people say you created the data
  - No Derivative Works – people can use your data but can only share it in unchanged form
  - ShareAlike – derivative works must be released under the same terms
  - Non-commercial – no commercial use allowed

Free your data

- CC licences restrict the possibilities of follow-on research. It is therefore recommended that a Creative Commons CC0 waiver is used instead.
- The CC0 waiver surrenders rights to the data as far as possible
- Good research practice means that people should cite your data if it is used
- The (work in progress) Creative Commons 4.0 licenses aim to be more data friendly than the current CC 3.0 license

Reasons not to publish

- Ethical / Legal Reasons
  - Unless previously agreed, people should not be identifiable from your data
  - Anonymising data may allow it to be published

- Licenses
  - Does the license for source data prevent you from publishing your data (e.g. use of CC-BY-SA data)
  - Is your work under a non-disclosure agreement (NDA)

Whose data is it anyway ?

- Chances are you do **not** own your research data
- Your contract may assign rights to everything you create as part of your research to your employer – including any data
- The data is probably owned by one of:
  - Your institution / employer
  - An industry partner
  - The funding body
- If you carry out a survey or interviews, the participants will hold the copyright on their input – unless you get them to transfer the rights to you!

Policies and Principles

- There may be policies and principles which state what should be done with your data
  - Institutional (e.g. QMUL)
  - Funder (e.g. EPSRC)
  - Publisher (e.g. IEEE)
- Policies and principles may cover:
  - Privacy – are you allowed to publish data ?
  - Publication – are you expected to publish data ?
  - Repositories – where should you publish data ?
  - Licences – who should be allowed to access data ?

EPSRC Principles

- The UK Engineering and Physical Sciences Research Council (EPSRC) states:
  - Data should be freely available with as few restrictions as possible
  - Data should remain accessible and usable for future research (10 years after last use!)
  - Metadata should be available to enable reuse
  - Results should say how to access the data
  - Users should acknowledge the sources of their data
  - Data management policies and plans should exist

http://tinyurl.com/993p6v6

Doing Data Management

- Easiest to start it at the start of a piece of work
- Do your work intending that it will be published
- Make it easy to publish your work at the end
- By 2015, QMUL should have a research data management system to store your PhD data!

Conclusions

- Data is fragile
  - computers break
  - media and formats become obsolete
- Without documentation, data becomes unusable
- Organising your data makes it more manageable
- Publish the data that validates your research

More Information ?

- Sound Data Management Training Wiki:
  - https://code.soundsoftware.ac.uk/projects/sodamat/wiki
- Vitae researcher development
  - http://www.vitae.ac.uk
  - Informed Researcher booklet
- Digital Curation Centre
  - http://www.dcc.ac.uk/