soundsoftware.ac.uk

Sustainable Data for Audio & Music Research

Steve Welburn

Centre for Digital Music

Queen Mary, University of London

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!)

The Lost Laptop Problem

» 2010 Ponemon Institute report for Intel re. US laptops

- 33% are lost in transit / while travelling
- 43% otherwise lost off-site
- 12% lost in the work-place
- 12% couldn't say
- Only ~4% recovered

Very similar results from 2011 European report!

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

http://tinyurl.com/876qza5

More ways to lose data...

» Running buggy code and overwriting your data

» Deleting a folder to salvage disk space

» Deleting the wrong file

» Losing a USB stick

» Forgetting to label DVDs

» Virus attack

» Disasters - fire, flood...

» Letting other people use your computer

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 ?

Where To Keep Your Data

- Just a working copy on your laptop
  - What if you run some buggy code and overwrite it ?
  - What if you lose the laptop ?
  - What if you break the laptop ?

High risk!

Where To Keep Your Data

- » WC + another copy on your laptop
  - – ~~What if you run some buggy code and overwrite it ?~~
  - – What if you lose the laptop ?
  - – What if you break the laptop ?

Better!

Provides a backup if you corrupt your data.

Where To Keep Your Data

» WC + a separate copy
(e.g. on another machine, on the 'net, on a removable drive)

  – ~~What if you run some buggy code and overwrite it ?~~
  – ~~What if you lose the laptop ?~~
  – ~~What if you break the laptop ?~~

Physically separate copies provide safe backups!

Don't keep your backup with your laptop as losing the laptop will then mean you've lost your backup!

Make sure you can find the backup if you need it!

Where To Keep Your Data

- Keeping copies of data in separate locations protects you from 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

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!

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 ?

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

- 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!
- 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, it's best to let someone else do it by placing your data in an archive which will deal with these issues for you.

Document

» Archiving data allows it to be accessed at a later date, but if someone looks at your data will they 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 ?
  - – License – how can it be used ?
  - – Audience – who might be interested ?
- » Metadata dœs not need to be structured, a README file explaining the file contents is sufficient.
- » Keeping documentation with the data means it is readily available

Organise

- File Names
  - Should be meaningful and brief
  - Should not depend on the folder structure
    - Files may be copied to different folders

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

Organise

- Folder Structures
  - A folder should contain either:
    - Subfolders
    - or a single type of file (e.g. code, data)

- If folders contain a single type of file, a general README can explain the content of each file in the folder

Publish

- Data can be published through a project web-site, but a data repository is a better solution as it should have a longer life
- Repositories may be:
  - Institutional i.e. location-specific
  - Thematic i.e. subject-specific
- Repositories are intended to provide long-term storage
- Data can be published in multiple repositories, but should have one identifier that is used to cite the data

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

Reasons not to publish

- Anonymisation
  - Unless previously agreed, people should not be identifiable from your data
- Ethical concerns
  - e.g. publishing bird song extracts live putting rare species at risk by revealing their location
- Licenses
  - Dœs the license for source data prevent you from publishing your data (e.g. use of CC-BY-SA data)

Where to publish ?

- » Institutional repository – if one exists!
- » Project or research group web-sites
- » Journal Supplementary Materials
  - – e.g. JASA, JNMR, CMJ
  - – Check T&Cs – 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!

Licensing Research Data

- If you don't supply a license, you reserve all rights to its use
- It is recommended that a Creative Commons CC0 waiver is used – this surrenders rights to the data as far as possible
- Copyright dœs not exist on factual data itself, only on the "creative" part of the data – e.g. the layout of a spreadsheet
- Attribution and Non-Commercial CC licenses may prevent people from using your data
- 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 licenses

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
  - Funder
  - Publisher
- 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 ?
  - Licenses – who should be allowed to access data ?

EPSRC Principles

- The UK Engineering and Physical Sciences Research Council (EPSRC http://tinyurl.com/993p6v6) 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

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