soundsoftware.ac.uk

Reusable Software and Reproducibility in Music Research

- Data

Mark Plumbley
(based on work by 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

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!

Local backups!

»We can keep local copies of files

- But what if the laptop is stolen ?

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

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

But I won't lose my laptop!

»So...

- What if you break the laptop ?
- What if the hard disk fails ?

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/98av3rd

Local Backups

»A working copy and a backup, stored on your laptop

- **Still High risk!**
- A separate backup is a better option!

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!

Student pleads for PhD laptop

...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.

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.

»So this is back where we started. Remember, to take care when you recover from a backup! Even better backup your current state before any recovery takes place!

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

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!

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 ?

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 does 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 Data

- 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 README can easily explain the contents of the files in a folder

Organise

- 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 is unreliable information as can change when files are moved
- Example:
  - Bad: piano.wav
  - Good: sjw_e12_20120829_piano.wav

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

» Releasing small datasets allows a larger corpus to be built from them.

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

- Does 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 – 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!

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 does 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) 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

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