Formatting, Organising and Storing Data

Scott Summers
UK Data Service
University of Essex
Cancer Research UK Data Sharing and Management
24th May 2017
Overview

• Looking after research data for the longer-term and protecting them from unwanted loss requires having good strategies in place for:

  • formatting
  • organising
  • securely storing
  • backing-up
  • transmitting / encrypting
  • and disposing of data

• Collaborative research brings additional challenges for the shared storage of, and access to, data
Stuff Happens!
Stuff happens: data inferno

- A fire destroyed a University of Southampton research centre resulting in significant damage to data storage facilities

- What if this was your office or your data?
- Source: BBC
Stuff happens: fieldwork nightmares

- “I’m sorry but we had to blow up your laptop.”

- “What….all my client case notes and testimony, writing, pictures, music and applications. Years of work. NO!!!!”

- Source: https://lilyasussman.com
Stuff happens: data theft

• What would happen if you lost your data?
• Imagine if you lost four years worth of research data – this nightmare situation happened to Billy Hinchen

Source: https://figshare.com/blog/The_stuff_of_nightmares_imagine_losing_all_your_research_data/121
Formatting, Organising and Storing Data
Versioning data

- Version control of files:
  - how many versions to keep? How long for?
  - it can be difficult to identify the correct version of a file if no standard naming practice is implemented
  - dates - 2016-10-20
  - spacing? - Interview Transcript 1
  - interview20%Transcript20%1
  - underscores - Interview_Transcript_1
  - major revisions vs minor revisions
  - 02-00
  - 02-01
  - FINAL vs FINAL? (DATA FINAL Presented FINAL)
Organising data

- Plan in advance how best to organise data
- Use a logical structure and ensure collaborators understand

Examples

- Hierarchical structure of files, grouped in folders, e.g. audio, transcripts and annotated transcripts
- Survey data: spreadsheet, SPSS, relational database
- Interview transcripts: individual well-named files
# Recommended file formats

<table>
<thead>
<tr>
<th>Documentation and scripts</th>
<th>Rich Text Format (.rtf)</th>
<th>plain text (.txt)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDF/UA, PDF/A or PDF (.pdf)</td>
<td>widely-used formats: MS Word (.doc/.docx), MS Excel (.xls/.xlsx)</td>
<td></td>
</tr>
<tr>
<td>XHTML or HTML (.xhtml, .htm)</td>
<td>XML marked-up text (.xml) according to an appropriate DTD or schema, e.g. XHMTL 1.0</td>
<td></td>
</tr>
<tr>
<td>OpenDocument Text (.odt)</td>
<td>Image data</td>
<td>TIFF 6.0 uncompressed (.tif)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Textual data</th>
<th>Rich Text Format (.rtf)</th>
<th>Hypertext Mark-up Language (.html)</th>
</tr>
</thead>
<tbody>
<tr>
<td>plain text, ASCII (.txt)</td>
<td>widely-used formats: MS Word (.doc/.docx)</td>
<td></td>
</tr>
<tr>
<td>eXtensible Mark-up Language (.xml) text according to an appropriate Document Type Definition (DTD) or schema</td>
<td>some software-specific formats: NUD*IST, NVivo and ATLAS.ti</td>
<td></td>
</tr>
</tbody>
</table>

| JPEG (.jpeg, .jpg, .jp2) if original created in this format |
| GIF (.gif) |
| TIFF other versions (.tif, .tiff) |
| RAW image format (.raw) |
| Photoshop files (.psd) |
| BMP (.bmp) |
| PNG (.png) |
| Adobe Portable Document Format (PDF/A, PDF) (.pdf) |

https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats
Quality control of data

• Quality control of data
  • Integral part throughout the research project
    • during data collection
    • data entry
    • data checking
  • Data collection and entry
    • calibration of instruments
    • taking multiple measurements
    • using standardised methods and protocols
  • Data checking
    • checking inputted correctly
    • checking data completeness
    • adding variable and value labels
Data storage

- Local storage
- Collaborative storage
- Cloud storage
- Data archives or repositories
Local data storage

- Internal hard drive / flash drive
- Note that all digital media are fallible
- Optical (CD, DVD and Blu-ray) and magnetic media (hard drives and tape) degrade over time
- Physical storage media become obsolete e.g. floppy disks

- Data files should be copied to new media every two-to-five years after they are first created
Collaborative storage

Sharing data between researchers

- Too often sent as insecure email attachments
- Physical media?
- Virtual Research Environments
  - MS SharePoint
  - Clinked
  - Huddle
  - Basecamp
- Cloud storage?
Cloud storage services

- Online or ‘cloud’ services are becoming increasingly popular
- Google Drive, DropBox, Microsoft OneDrive and iCloud

Benefits:
- very convenient
- accessible anywhere
- good protection if working in the field?
- background file syncing
- mirrors files
- mobile apps available

But,
- these are not necessarily secure
- potential DPA issues
- not necessarily permanent
- intellectual property right concerns?
- limited storage?
Cloud storage services

• Perhaps more secure options?

Mega.nz  SpiderOak  Tresorit

• Cloud data storage should be avoided for high-risk information such as files that contain personal or sensitive information, or information that has a very high intellectual property value.
File sharing – data archive or repository

• A repository acts as more of a ‘final destination’ for data

• UK Data Service has its own data repository called ‘ReShare’, for social science data of any kind

• [http://reshare.ukdataservice.ac.uk/](http://reshare.ukdataservice.ac.uk/)
Backing-up Data
Backing-up data

- It is not a case of *if* you will lose data, but *when* you will lose data!
- Keep additional backup copies and protect against: software failure, hardware failure, malicious attacks and natural disasters
- **Would your data survive a disaster?**
Digital back-up strategy

Consider

• **What’s backed-up?** - all, some or just the bits you change?
• **Where?** - original copy, external local and remote copies
• **What media?** - DVD, external hard drive, USB, Cloud?
• **How often?** - hourly, daily, weekly? Automate the process?
• **What method/software?** - duplicating, syncing or mirroring?
• **For how long is it kept?** - data retention policies that might apply?
• **Verify and recover** - never assume, regularly test and restore

Backing-up need not be expensive

• 1Tb external drives are around £50, with back-up software

Also consider non-digital storage too!
Verification and integrity checks

• Ensure that your backup method is working as intended
• Automated services - check
• Be wary when using sync tools in particular
  • mirror in the wrong direction or using the wrong method, and you could lose new files completely
• You can use checksums to verify the integrity of a backup
• Also useful when transferring files
• Checksum somewhat like a files’ fingerprint
• …but changes when the file changes
Checksums

- Each time you run a checksum a number string is created for each file

- Even if one byte of data has been altered or corrupted that string will change

- Therefore, if the checksums before and after backing up a data file match, then you can be sure that the data have not altered during this process

- A free software tool for computing MD5 checksums is MD5summer for windows
Data Security
Data security

Protect data from unauthorised
• Access
• Use
• Change
• Disclosure
• Destruction

Who knows who is watching, listening or attempting to access your data…

GCHQ IS NOW FOLLOWING YOU ON TWITTER!... AND FACEBOOK... AND EMAIL... AND...
Data security strategy

• Control access to computers:
  • use passwords and lock your machine when away from it
  • run up-to-date anti-virus and firewall protection
  • power surge protection
  • utilise encryption
    • on all devices: desktops, laptops, memory sticks and mobile devices
    • at all locations: work, home and travel
  • restrict access to sensitive materials e.g. consent forms and patient records
  • personal data need more protection – always keep them separate and secure

• Control physical access to buildings, rooms and filing cabinets

• Properly dispose of data and equipment once your project is finished
Passwords

• Strong passwords are crucial

• Avoid using weak or easy to guess passwords and reusing passwords

• Consider password managers, complex passwords or stringing words together to create stronger passwords

• But, remember that you need to be able to remember the passwords!

• Why does this matter?

• No matter how good the encryption is that you use if you use a weak password the encryption will offer little protection

https://howsecureismypassword.net (*Never use real passwords)
Password Security

HOW SECURE IS MY PASSWORD?

―

“Password”

Your password would be cracked instantly.

Why not try Dashlane to create and remember stronger passwords? It's free!

UK Data Service
Password Security

HOW SECURE IS MY PASSWORD?

It would take a computer about

27 UNDECILLION YEARS

to crack your password

Dashlane can help you remember all of your secure passwords - and it's free!
Encryption

- Encryption is the process of encoding digital information in such a way that only authorised parties can view it.

- **Always** encrypt personal or sensitive data
  - = anything you would not send on a postcard
  - e.g. moving files, such as interview transcripts
  - e.g. storing files to shared areas or insecure devices

- Basic principles
  - applies an algorithm that makes a file unreadable
  - needs a ‘key’ of some kind (passphrase or/and file) to decrypt

- The UK Data Service recommends Pretty Good Privacy (PGP)
  - more complicated than just a password, but much more secure
  - involves use of multiple public and private keys
Encryption software

Encryption software can be easy to use and enables users to:
• encrypt hard drives, partitions, files and folders
• encrypt portable storage devices such as USB flash drives

VeraCrypt

BitLocker

Axcrypt

FileVault2
Data disposal

- When you delete a file from a hard drive, it is likely to still be retrievable (even after emptying the recycle bin).
- Even reformatting a hard drive is **not** sufficient.
- Files need to be overwritten multiple times with random data for best chances of removal.
- The **only** sure way to ensure data is irretrievable is to physically destroy the drive (using an approved secure destruction facility).
Questions

Contact details:

Collections Development and Producer Relations team
UK Data Service
University of Essex
ukdataservice.ac.uk/help/get-in-touch