Backup strategy

source : UK Data Archive[1]/ ISSC

Back-up procedures depend on local circumstances, the perceived value of the data and the levels of risk considered appropriate. A backup is a copy of (a grouping of) files at a certain point in time, in order to address the risk of losing the primary data due to human error, theft, or accidents. It is advised to backup critical data or frequently used data at least daily, preferably by using an automated backup procedure. Another best practice is to store this backup at an alternate location. Besides preventing the risk of losing data, one can create an archive of (grouping of) files on alternate media for long-term storage reasons, in order to make data transport easier, or to free up directly accessible storage space.

Questions to ask in order to design your strategy :

1. Is there any backup provision already in place?
2. Which files / systems to back up?
3. How often? After each change to a data file or at regular intervals ?
4. Where should I store my backups?
5. How should I organize my backups?
6. How about backing up personal data?
7. Are there any tools available?
8. Anything else to think about?

Some answers

The Leiden University servers have a nightly backup. For details see : http://media.leidenuniv.nl/legacy/minimale-maatregelen-informatiebeveiliging.pdf [Dutch].

Back-up files can be kept on a networked hard drive (most convenient) or stored offline on media such as recordable CD/DVD (small datasets), removable hard drive or magnetic tape (large quantities of data). Make sure your removable media are well labeled (content, date/time) and organized so they are easy to locate and physically accessible.

Where data contain personal information, care should be taken to only create the minimal number of copies needed, e.g. a master file and one back-up copy (think of deletion of data!).

Back-up files should be verified and validated regularly, either by fully restoring them to another location and comparing them with the originals or by checking back-up copies for completeness and integrity, for example by checking the MD5 checksum value (quantitative data) or file size (see below).

[1] Based on : http://ukdataservice.ac.uk/manage-data/store/backup.aspx ; *Managing and sharing research data : a guide to good practice* / Louise Corti, Veerle Van den Eynden, et al., Los Angeles : Sage, 2014.

Recommendations from the UK Data Archive :

- ✓ store data uncompressed in non-proprietary or **open standard formats** for long-term software readability ('preferred formats'[2])
- ✓ **copy or migrate** data files to new media between two and five years after they were first created, since both optical and magnetic media are subject to physical degradation
- ✓ use a storage **strategy**, even for a short-term project, with two different forms of storage, e.g. on hard drive and on CD
- ✓ create **digital versions** of paper documentation in PDF/A format for long-term preservation and storage

System or file backup ?

Full back- ups consist of making a copy of all relevant files, often the complete contents of a PC or share. Restoring files from such a backup generally only requires the latest backup to be available.

For **differential** back-ups, a complete back-up is made first, and then back-ups are made of files changed or created since the first full back-up and not just since the last partial back-up. Fixed media, such as hard drives, are recommended for this method. Restoring files from differential backups require both the full backup and the latest differential. The retrieval time is longer but a differential back-up requires much less storage space.

Incremental back-ups consist of first making a copy of all relevant files, often the complete contents of a PC, then making incremental back-ups of the files which have altered since the last back-up. Removable media (CD/DVD) are recommended for this procedure. This method requires even less storage space, but retrieval time and complexity increases dramatically as all partial backups need to be accessed as well as the full backup.

Whichever method is used, it is best not to overwrite old back-ups with new.
UK Data Archive :

Tools for back-up : http://en.wikipedia.org/wiki/List_of_backup_software

Check the integrity of your data regularly with a checksum checker (source : UKDA[3])

A checksum is like a fingerprint of a file and can be used to verify whether two files are identical.
Each time you run a checksum a number is created for each file. Even if one byte of data has been altered or corrupted that string will change. So, if the checksums before and after copying or backing up a data file match, then you can be sure that the data have not altered during this process.

The checksum values can be easily stored and shared, and compared against the value for a non-identical file. For example, a checksum list copied onto a USB stick will allow you to detect problems early.

If you find any differences between the two checksums, then you should transfer your data onto another storage device immediately.

There are free tools available for all sorts of systems.

[2] See for instance : http://dans.knaw.nl/en/deposit/information-about-depositing-data/DANSpreferredformatsUK.pdf
[3] Ibid., p. 95