Spreadsheet vs Database

Avoiding Data Disasters

Anne Pajon, CRUK-CI

File Management

Avoiding Data Disasters

Anne Pajon, CRUK-CI

Use descriptive and informative **file names**

File Names ... Best Practices

- Do not name all your data files '**`data.xls`**'
- Include any information that will allow you to distinguish your files from one another
  - Project or experiment name or acronym
  - Location/spatial coordinates
  - Researcher name/initials
  - Date or date range of experiment
  - Type of data
  - Conditions
  - Version number of file
  - Three-letter file extension for application-specific files
- Choose a consistent naming scheme across all your files
- Include in the directory a '**`README.txt`**' file that explains your naming format along with any abbreviations or codes you have used

File Names ... Other Tips (1)

- **Avoid special characters** such as ~ ! @ # $ % ^ & * ( ) ` ; < > ? , [ ] { } ' " and |
- **Use short file names**, long ones do not work well with all types of software
- A **good format for date** designations is **`YYYYMMDD or YYMMDD`**
  - All of your files stay in chronological order, even over the span of many years
- Use **leading zeros** for clarity and to make sure files sort in sequential order
  - For example, use "001, 002, ...010, 011 ... 100, 101, etc." instead of "1, 2, ...10, 11 ... 100, 101, etc."

File Names ... Other Tips (2)

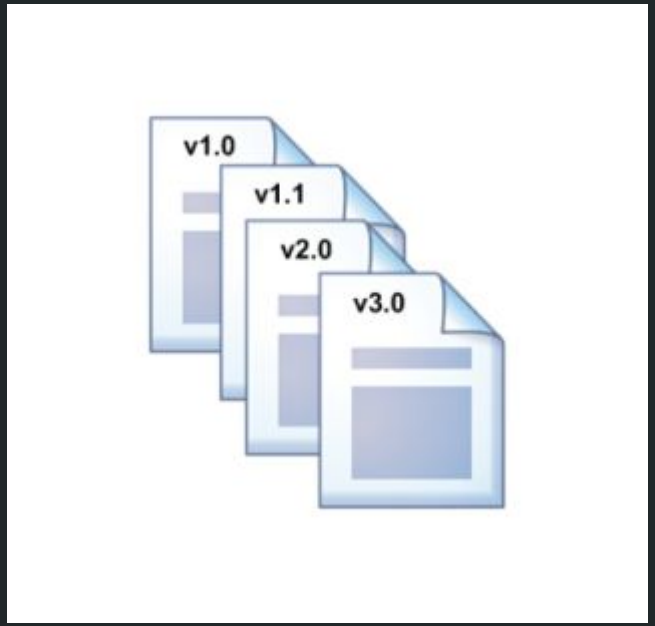
- **Do not use spaces**. Some software will not recognize file names with spaces, and file names with spaces must be enclosed in quotes when using the command line. Other options include:
  - Underscores, e.g. `file_name.xxx`
  - Dashes, e.g. `file-name.xxx`
  - No separation, e.g. `filename.xxx`
  - Camel case, where the first letter of each section of text is capitalized, e.g. `FileName.xxx`

Choose **file formats** that will ensure long-term access

File Formats ... Best Practices

- Save data in a **non-proprietary** (open) file format when possible
  - Usable on diverse platforms and by multiple applications
  - Export your data as tab separated file
- Unencrypted
- Uncompressed
- In common usage by the research community
- Preferred formats
  - .csv, .txt

Track different **versions** of your documents



Data versioning

Versioning refers to saving new copies of your files when you make changes allowing you to reverse or roll back those changes or retrieve specific versions of your files later

- Simple file versioning
- Simple software options
- Advanced software options

Simple File versioning

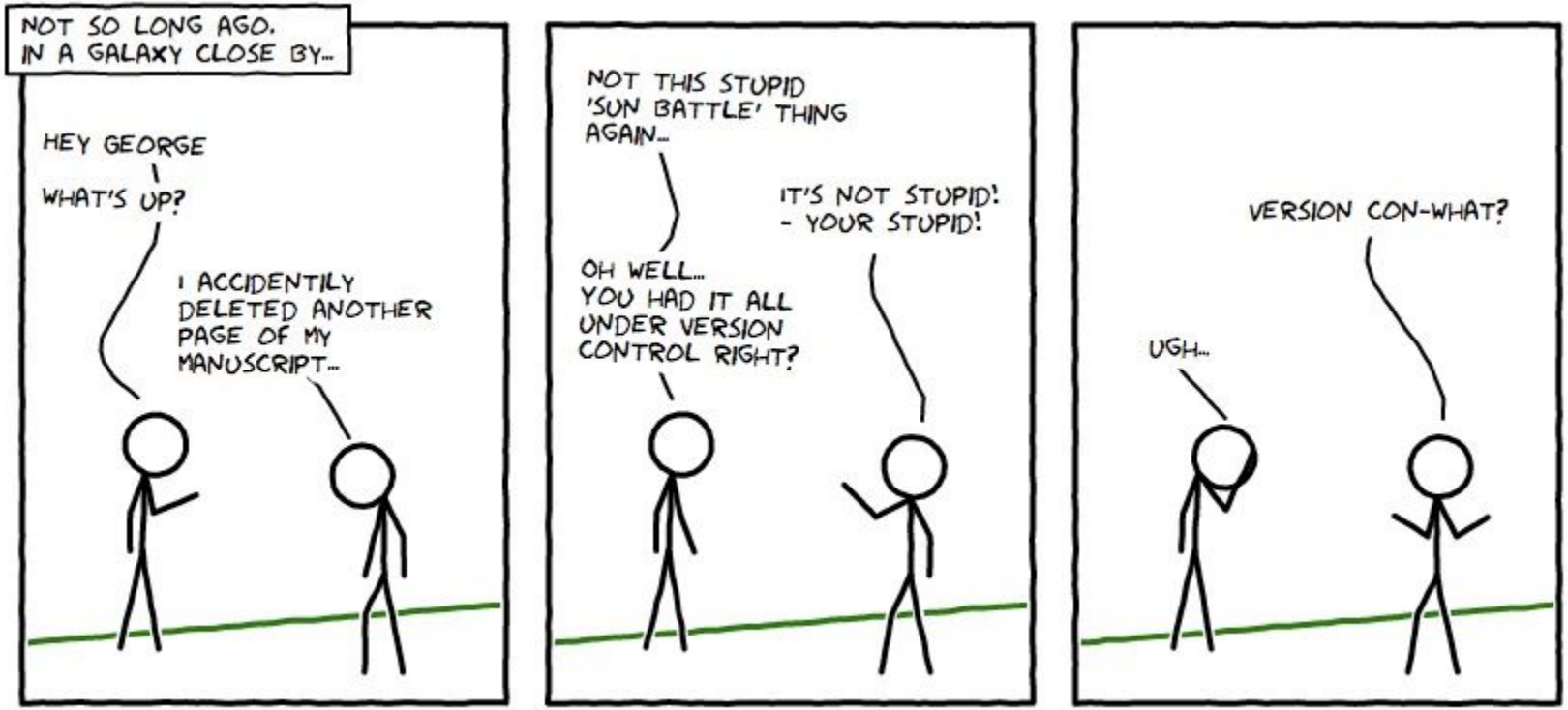
- Manually save new versions when you make significant changes
  - Include a version number, e.g. "v1," "v2," or "v2.1" into file names
- This works well, only if...
  - No need to keep lots of different versions
  - Only one person working on these files
  - Always access these files from one location

Simple Software Options

- Use Google Drive's word processing, spreadsheet and presentation
  - Any time you edit files, new versions are saved as you go
  - Version information includes who was editing the file and when the new version was created

- Dropbox free version keeps track of all of the changes you make for 30 days
  - The paid Packrat version keeps track of every change you ever make to the files in your Dropbox

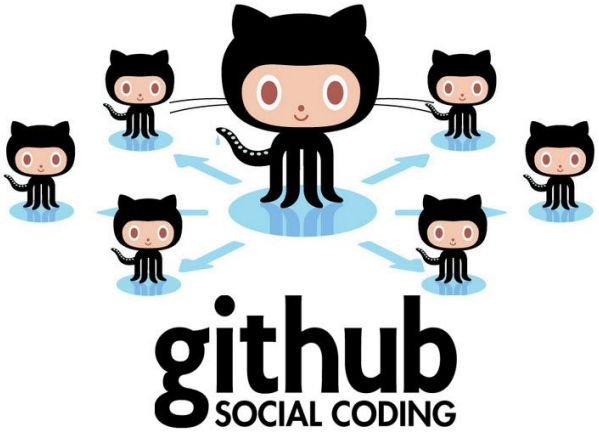
Advanced Software Options



Advanced Software Options

Version control systems like subversion and git are frequently used for groups writing software and code, but can be used for any kind of files or projects. Many people share their git repositories on GitHub.

TODO. What is version control?



GitHub Desktop

Overview | Release Notes | Help

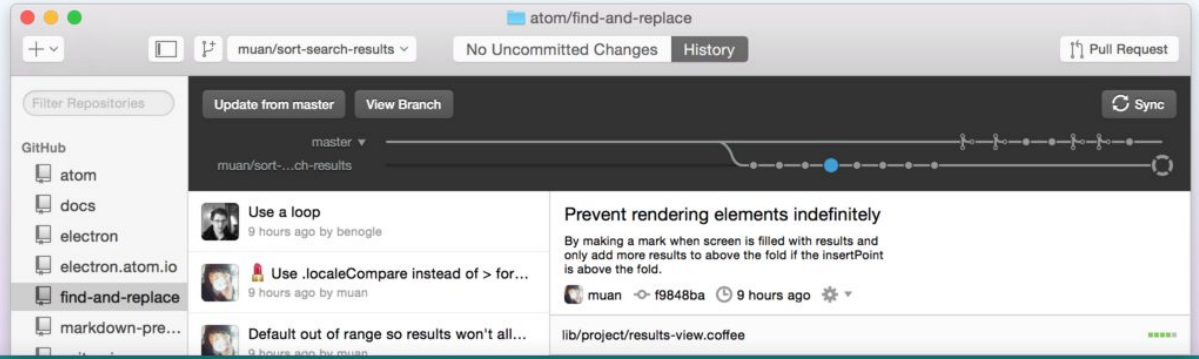
Simple collaboration from your desktop

GitHub Desktop is a seamless way to contribute to projects on **GitHub** and **GitHub Enterprise**.

Available for Mac and Windows

Download GitHub Desktop
OS X 10.9 or later

By clicking the Download button you agree to the End-User License Agreement



Your GitHub workflow in one native app

Clone repositories → Create branches → Commit changes → Share code

Example: Training material on github

Reference

- Data best practices STANFORD UNIVERSITY LIBRARIES
  - http://library.stanford.edu/research/data-management-services/data-best-practices

- Excel vs Databases SCHOOL OF DATA
  - http://schoolofdata.org/2013/11/07/sql-databases-vs-excel/