Tools for version control of research data

Raf Guns
raf.guns@uantwerpen.be

University of Antwerp

Electronic lab notebooks

Mercurial

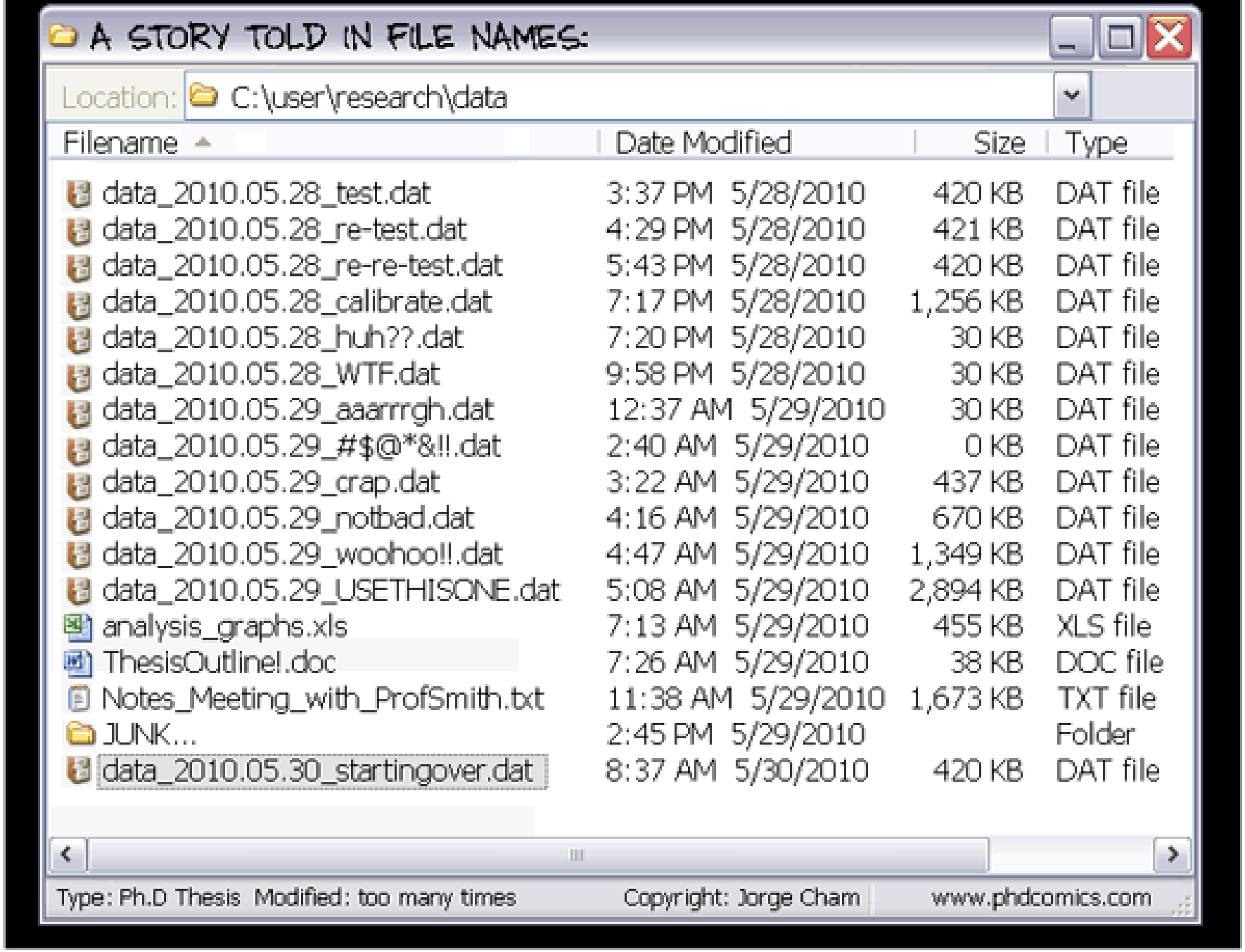
Version control

Git

Subversion

Revision control

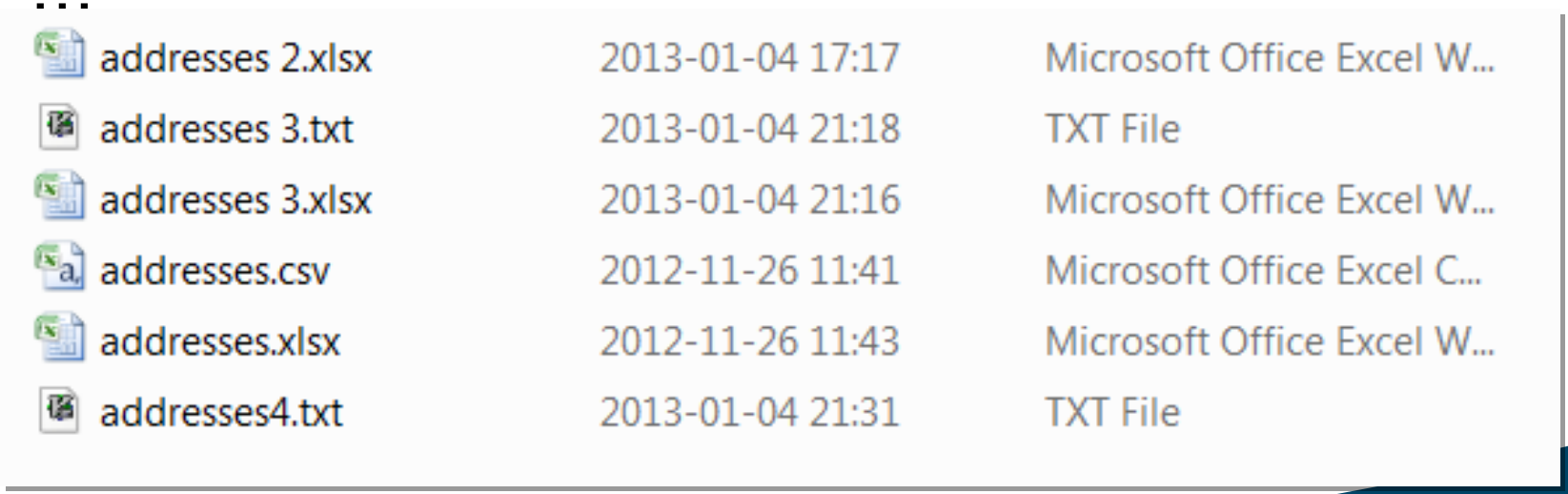
Github



http://www.phdcomics.com/comics/archive.php?comicid=1323

Research data: not static

- Data cleaning
- Correcting errors
- Multiple data sources
- Data reuse
- Conversion between file formats
- …



Version control

“the **management of changes** to documents, computer programs, large web sites, and other collections of information”

(Wikipedia)

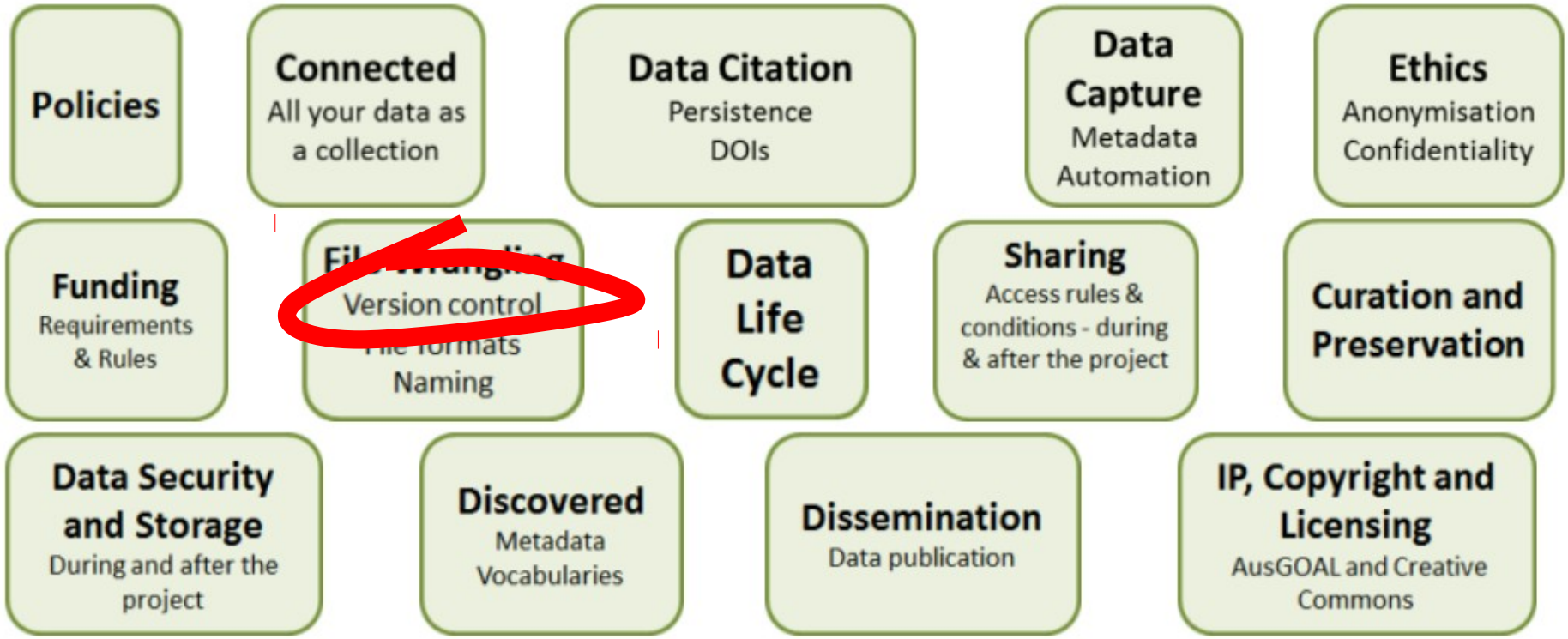
Why version control?

- **Revert** to previous versions
- Find out **what is different** between two versions
- Find out **what has changed** in a specific time period
- Manage **multiple versions**
- Work with **multiple people** on the same data
- **Transparency** and **integrity**

  http://www.scfbm.org/content/8/1/7

☞ *Version control is an integral part of*

Research data management



www.ands.org.au/resource/data-management-planning.html

Tools for version control

Built-in

~~To Whom it May Concern~~Dear Mr. Powell:

Thank you for taking the time to meet with me last Thursday about the Sales Associate position. I enjoyed meeting with you and touring the facility. I was very impressed with the layout of the showroom and with the competence of the staff at ~~your company~~Quality Furnishings. I would love the chance to work in such a productive and ~~very~~ supportive atmosphere.

As we talked about in our meeting, my many years of sales experience, both in commissioned floor sales and in the role of Sales Supervisor, would greatly benefit Quality Furnishings. In that time, I have learned many techniques that would ~~drive~~ increase sales and drive customer satisfaction ratings at Quality Furnishings.

p://www.gcflearnfree.org/word2010/20.2

Dropbox

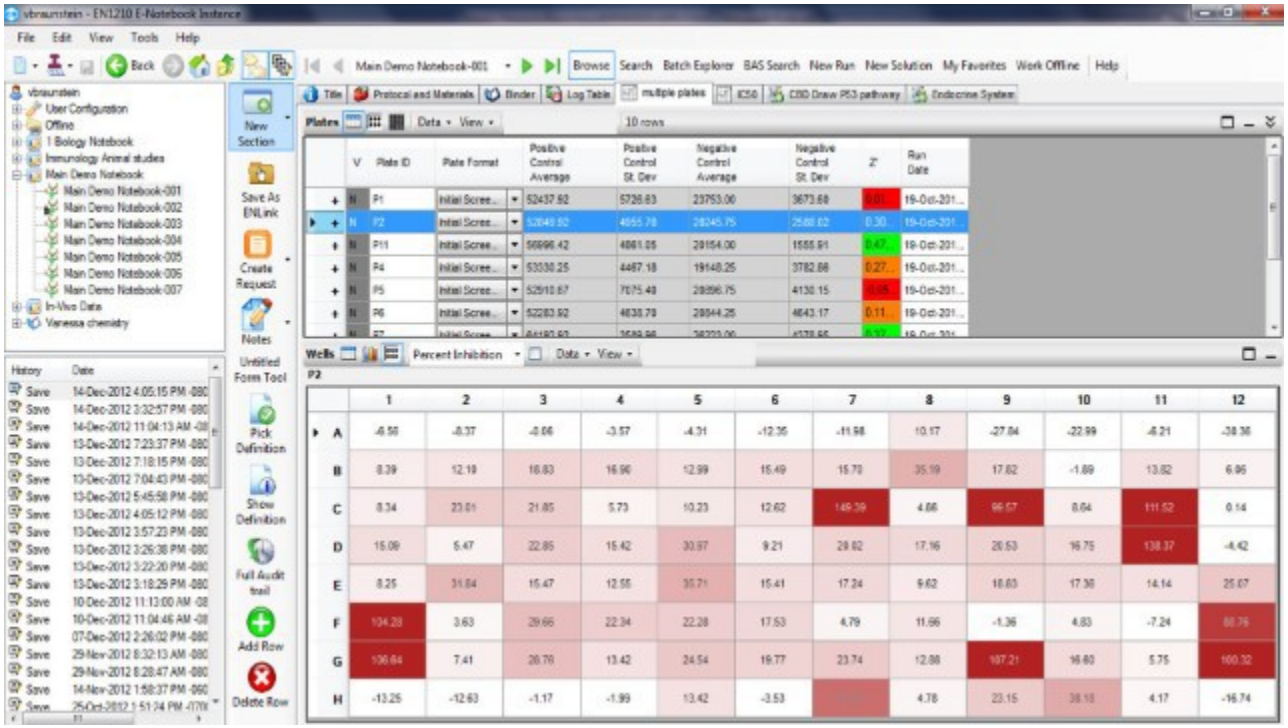
Version history of 'Richard Thompson article BB.md'

Dropbox keeps a snapshot every time you save a file. You can preview and restore 'Richard Thompson article BB.md' by choosing one of the versions below:

Version 18 (current)	Edited by Glenn Fleishman (Mininum-Minival)	4/12/2012 5:10 PM	11.71 KB
Version 17	Edited by Glenn Fleishman (Mininum-Minival)	4/12/2012 5:07 PM	11.57 KB
Version 16	Edited by Glenn Fleishman (Mininum-Minival)	4/12/2012 5:02 PM	11.16 KB
Version 15	Edited by Glenn Fleishman (Mininum-Minival)	4/12/2012 4:35 PM	9.39 KB
Version 14	Edited by Glenn Fleishman (Mininum-Minival)	4/12/2012 4:31 PM	9.62 KB
Version 13	Edited by Glenn Fleishman (Mininum-Minival)	4/12/2012 4:28 PM	9.57 KB
Version 12	Edited by Glenn Fleishman (Mininum-Minival)	4/12/2012 4:27 PM	9.47 KB
Version 11	Edited by Glenn Fleishman (Mininum-Minival)	4/12/2012 4:23 PM	9.28 KB
Version 10	Edited by Glenn Fleishman (Mininum-Minival)	4/12/2012 4:23 PM	9.27 KB

Electronic lab notebooks (ELNs)

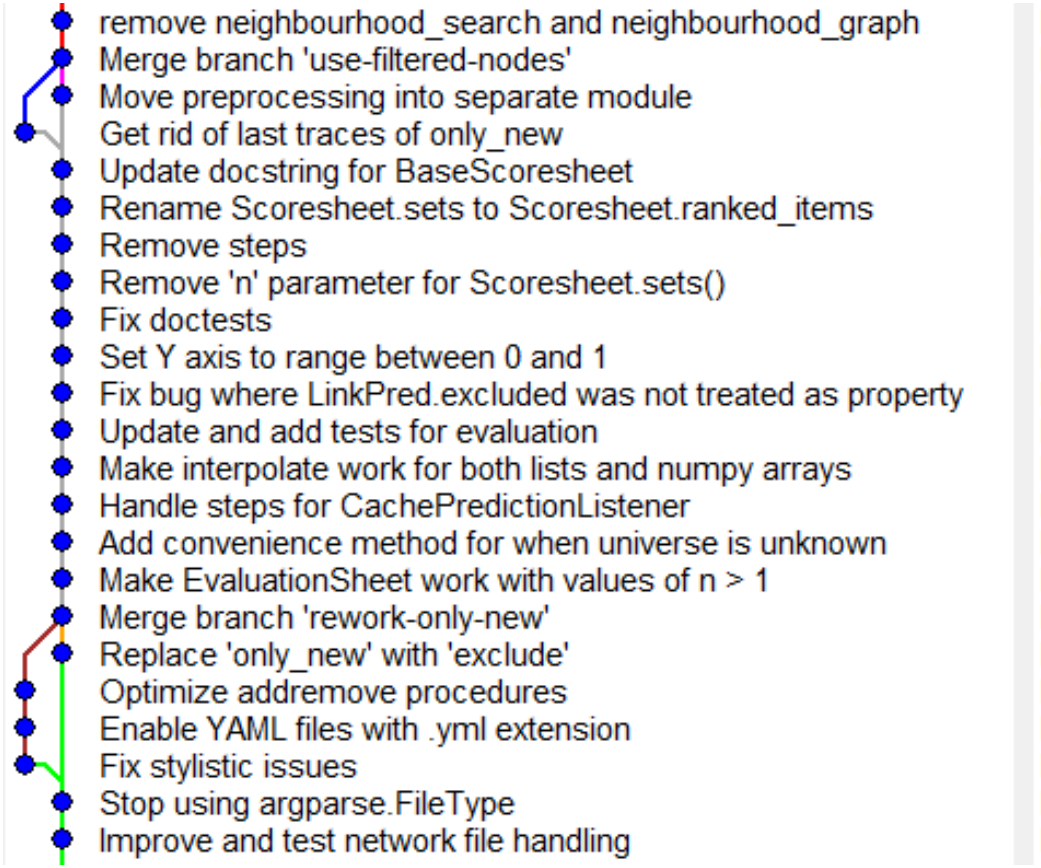
- Replacement of paper lab notebook
- Integrate text, figures, data, calculations…



Version control software: Git

- Alternatives: Mercurial, SVN, CVS, Perforce, Bazaar, Arch…
- We pick Git because
  - probably most popular nowadays
  - Github (www.github.com)
- Open source, free of charge
- Command-line, several GUIs available (gitk, Github for Mac, Github for Windows…)

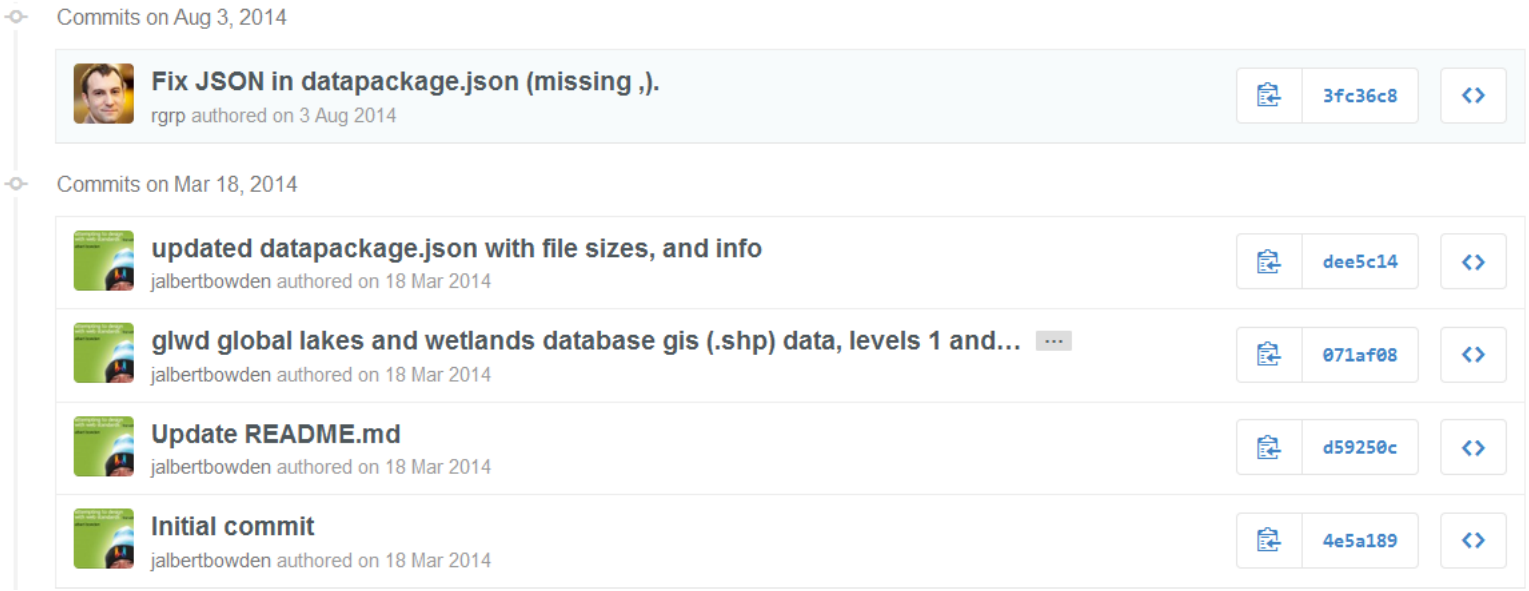
Git history



remove neighbourhood_search and neighbourhood_graph	Raf Guns <rafguns@gmail.com>	2014-12-04 15:03:00
Merge branch 'use-filtered-nodes'	Raf Guns <rafguns@gmail.com>	2014-07-29 13:46:22
Move preprocessing into separate module	Raf Guns <rafguns@gmail.com>	2014-07-29 13:06:23
Get rid of last traces of only_new	Raf Guns <rafguns@gmail.com>	2014-07-29 13:32:31
Update docstring for BaseScoresheet	Raf Guns <rafguns@gmail.com>	2014-07-10 16:15:42
Rename Scoresheet.sets to Scoresheet.ranked_items	Raf Guns <rafguns@gmail.com>	2014-07-10 15:28:14
Remove steps	Raf Guns <rafguns@gmail.com>	2014-07-10 12:03:18
Remove 'n' parameter for Scoresheet.sets()	Raf Guns <rafguns@gmail.com>	2014-07-10 10:34:36
Fix doctests	Raf Guns <rafguns@gmail.com>	2014-07-09 15:55:14
Set Y axis to range between 0 and 1	Raf Guns <rafguns@gmail.com>	2014-06-30 14:47:00
Fix bug where LinkPred.excluded was not treated as property	Raf Guns <rafguns@gmail.com>	2014-06-30 14:06:01
Update and add tests for evaluation	Raf Guns <rafguns@gmail.com>	2014-02-26 14:51:34
Make interpolate work for both lists and numpy arrays	Raf Guns <rafguns@gmail.com>	2014-01-13 22:42:55
Handle steps for CachePredictionListener	Raf Guns <rafguns@gmail.com>	2014-01-13 22:27:36
Add convenience method for when universe is unknown	Raf Guns <rafguns@gmail.com>	2014-01-13 17:48:04
Make EvaluationSheet work with values of n > 1	Raf Guns <rafguns@gmail.com>	2014-01-13 17:20:21
Merge branch 'rework-only-new'	Raf Guns <rafguns@gmail.com>	2014-06-18 10:58:47
Replace 'only_new' with 'exclude'	Raf Guns <rafguns@gmail.com>	2013-12-17 14:29:01
Optimize addremove procedures	Raf Guns <rafguns@gmail.com>	2014-06-16 11:39:01
Enable YAML files with .yml extension	Raf Guns <rafguns@gmail.com>	2014-01-13 13:49:28
Fix stylistic issues	Raf Guns <rafguns@gmail.com>	2014-01-10 14:53:38
Stop using argparse.FileType	Raf Guns <rafguns@gmail.com>	2013-12-12 11:39:43
Improve and test network file handling	Raf Guns <rafguns@gmail.com>	2013-12-12 11:39:06

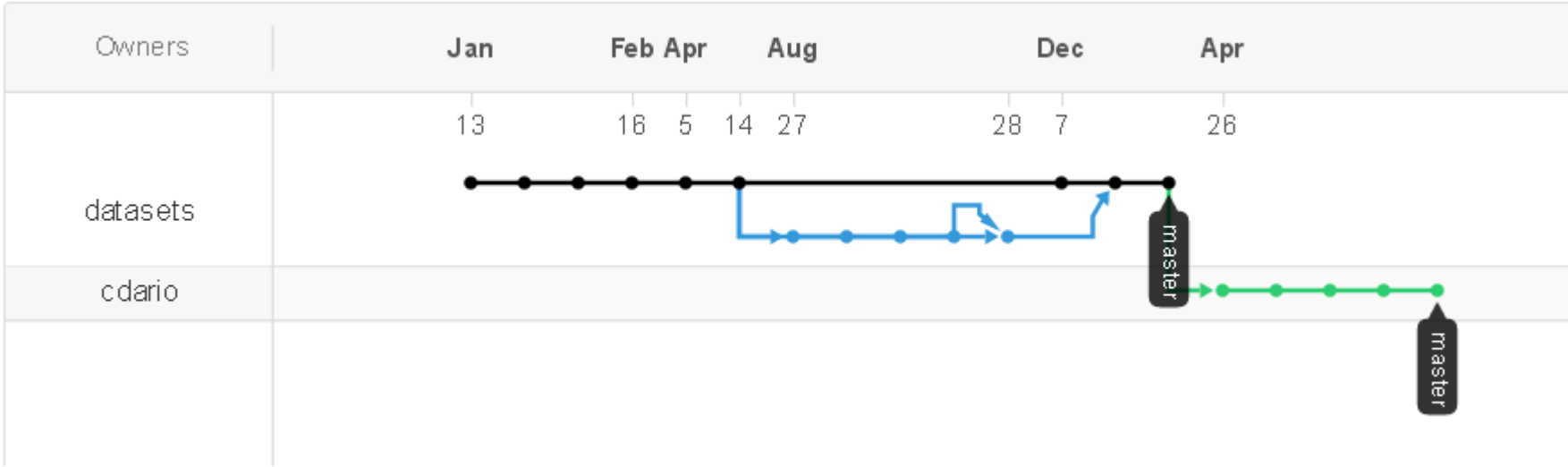
Data set history

https://github.com/datasets/glwd



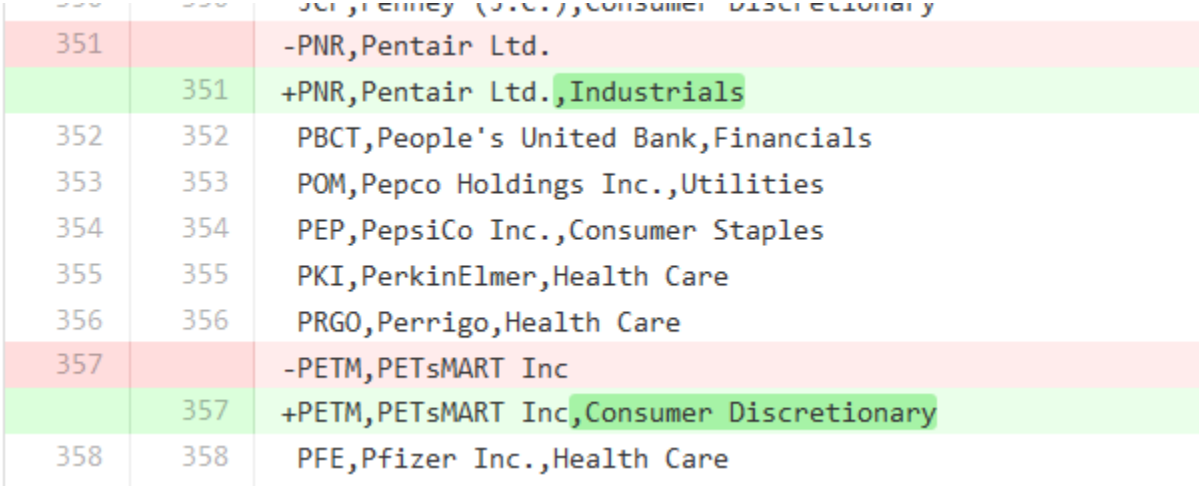
Data set branches

https://github.com/datasets/country-list



Data set ‘diff’

https://github.com/datasets/s-and-p-500-companies



Git for research data

Made for software development:

- many small files (e.g. Github has limitation of 100MB)
- text-based formats (e.g. CSV, XML, JSON…)

Some research data sets are:

- much larger
- ‘binary’ formats (non-readable without special software)

Solutions: git-annex, git-lfs…

Dat http://dat-data.com

In development, alpha software!

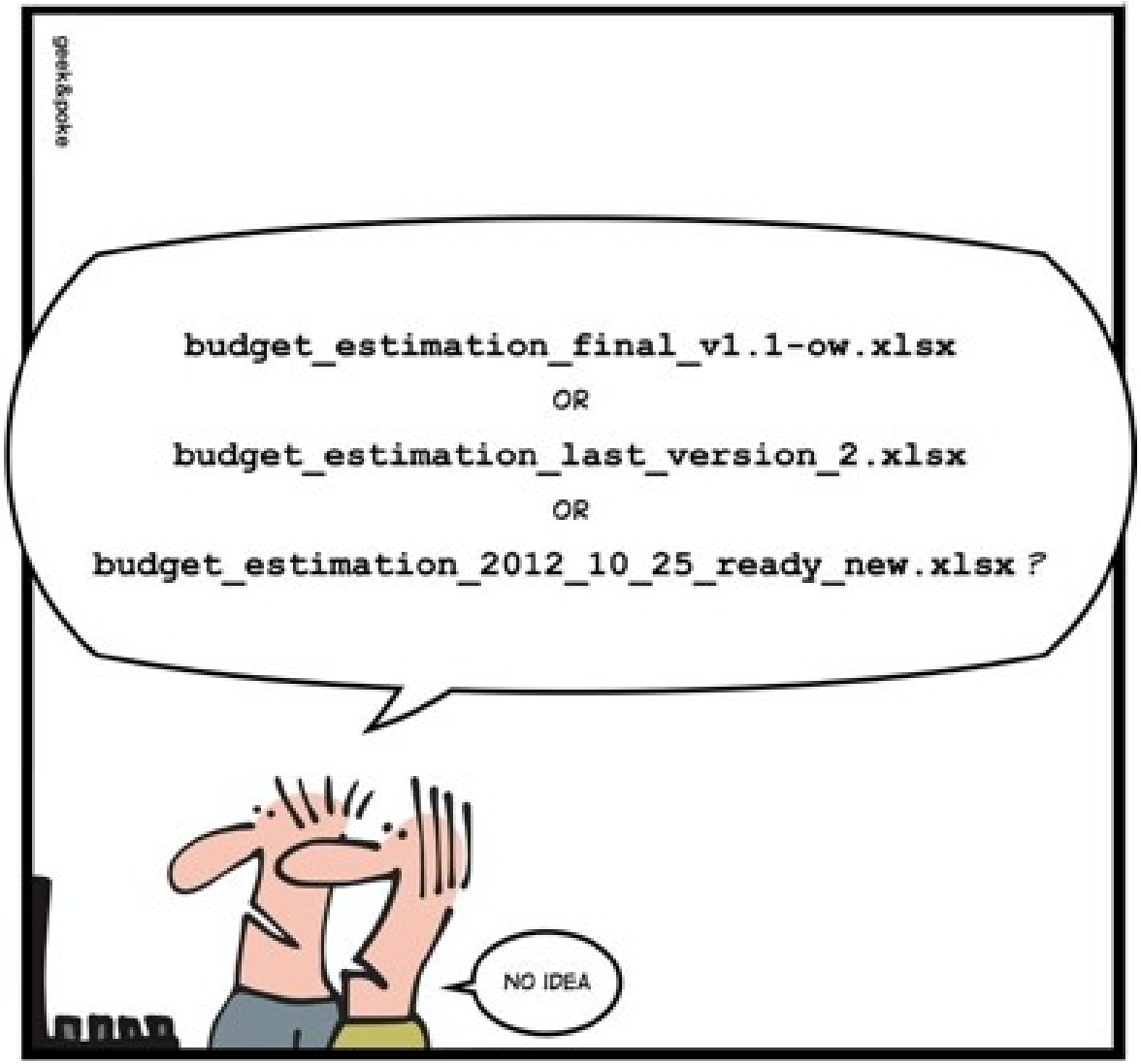
“real time replication and versioning for data sets”

Two kinds of data:

- tabular: can be expressed in table
- blobs: unstructured and/or large (e.g., images)

Automatically translates between different formats (e.g. JSON – XML)

Thank you!



//www.datamation.com/news/tech-comics-version-control-1.html