

# Tools for version control of research data

Raf Guns

[raf.guns@uantwerpen.be](mailto:raf.guns@uantwerpen.be)

Electronic lab  
notebooks

Mercuria  
l

Version control

Git

Subversi  
on

Revision control

Github



## A STORY TOLD IN FILE NAMES:

Location: C:\user\research\data

Filename	Date Modified	Size	Type
data_2010.05.28_test.dat	3:37 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
data_2010.05.29_#*\$@*&!!.dat	2:40 AM 5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM 5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file







Type: Ph.D Thesis Modified: too many times

Copyright: Jorge Cham

www.phdcomics.com

# Research data: not static

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

 addresses 2.xlsx	2013-01-04 17:17	Microsoft Office Excel W...
 addresses 3.txt	2013-01-04 21:18	TXT File
 addresses 3.xlsx	2013-01-04 21:16	Microsoft Office Excel W...
 addresses.csv	2012-11-26 11:41	Microsoft Office Excel C...
 addresses.xlsx	2012-11-26 11:43	Microsoft Office Excel W...
 addresses4.txt	2013-01-04 21:31	TXT File

# 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



# 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.

# 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:

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

The screenshot displays the Vbrunstein EN1210 E-Notebook Instance software interface. The main window shows a 'Plates' table with the following data:

V	Plate ID	Plate Format	Positive Control Average	Positive Control St. Dev	Negative Control Average	Negative Control St. Dev	Z'	Run Date
N	P1	Initial Screen...	52437.52	5726.83	23753.00	3673.68	0.01	16-Oct-201...
N	P2	Initial Screen...	52040.92	4955.78	28245.75	2580.82	0.30	16-Oct-201...
N	P11	Initial Screen...	58996.42	4881.05	28154.00	1555.91	0.47	16-Oct-201...
N	P4	Initial Screen...	53030.25	4487.18	19148.25	3782.88	0.22	16-Oct-201...
N	P5	Initial Screen...	52910.67	7175.48	28286.75	4130.15	0.08	16-Oct-201...
N	P6	Initial Screen...	52283.92	4830.79	28944.25	4643.17	0.11	16-Oct-201...
N	P2	Initial Screen...	52240.82	5628.88	36253.00	4738.65	0.22	16-Oct-201...

Below the 'Plates' table, the 'Wells' table for plate P2 is shown, displaying a grid of data for columns 1-12 and rows A-H. The data is color-coded, with red indicating high values and green indicating low values.

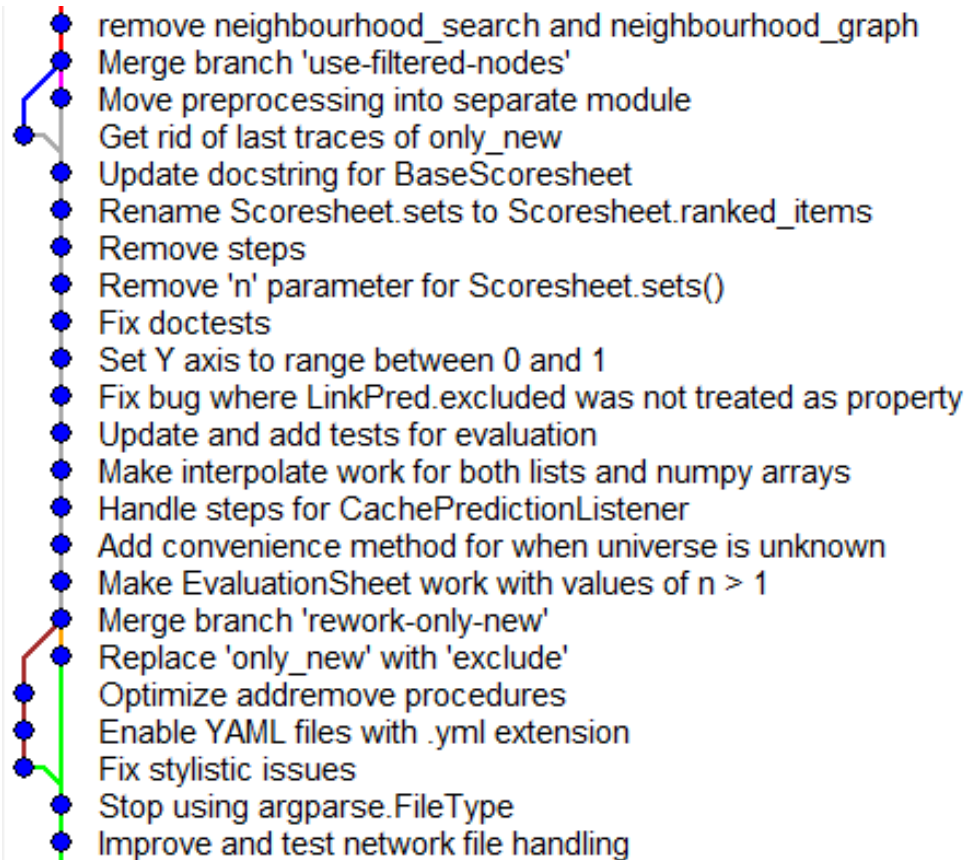
	1	2	3	4	5	6	7	8	9	10	11	12
A	-6.56	-8.37	-8.66	-3.57	-4.31	-12.35	-11.98	10.17	-27.84	-22.99	-6.21	-38.36
B	8.39	12.19	16.83	16.90	12.99	15.49	15.70	35.19	17.62	-1.89	13.82	6.96
C	8.34	23.61	21.85	5.73	10.23	12.62	146.59	4.66	99.57	8.04	111.52	0.14
D	15.09	5.47	22.85	15.42	30.97	9.21	28.82	17.16	26.53	16.75	138.37	-4.42
E	8.25	31.64	15.47	12.55	35.71	15.41	17.24	9.62	16.63	17.38	14.14	25.67
F	134.23	3.63	29.65	22.34	22.28	17.53	4.79	11.96	-1.36	4.83	-7.24	81.75
G	106.64	7.41	26.70	13.42	24.54	16.77	23.74	12.88	187.21	16.80	5.75	160.32
H	-12.25	-12.63	-1.17	-1.99	13.42	-3.53		4.78	23.15	38.13	4.17	-16.74

# Version control software: Git

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



# Git history



Raf Guns <rafguns@gmail.com>	2014-12-04 15:03:00
Raf Guns <rafguns@gmail.com>	2014-07-29 13:46:22
Raf Guns <rafguns@gmail.com>	2014-07-29 13:06:23
Raf Guns <rafguns@gmail.com>	2014-07-29 13:32:31
Raf Guns <rafguns@gmail.com>	2014-07-10 16:15:42
Raf Guns <rafguns@gmail.com>	2014-07-10 15:28:14
Raf Guns <rafguns@gmail.com>	2014-07-10 12:03:18
Raf Guns <rafguns@gmail.com>	2014-07-10 10:34:36
Raf Guns <rafguns@gmail.com>	2014-07-09 15:55:14
Raf Guns <rafguns@gmail.com>	2014-06-30 14:47:00
Raf Guns <rafguns@gmail.com>	2014-06-30 14:06:01
Raf Guns <rafguns@gmail.com>	2014-02-26 14:51:34
Raf Guns <rafguns@gmail.com>	2014-01-13 22:42:55
Raf Guns <rafguns@gmail.com>	2014-01-13 22:27:36
Raf Guns <rafguns@gmail.com>	2014-01-13 17:48:04
Raf Guns <rafguns@gmail.com>	2014-01-13 17:20:21
Raf Guns <rafguns@gmail.com>	2014-06-18 10:58:47
Raf Guns <rafguns@gmail.com>	2013-12-17 14:29:01
Raf Guns <rafguns@gmail.com>	2014-06-16 11:39:01
Raf Guns <rafguns@gmail.com>	2014-01-13 13:49:28
Raf Guns <rafguns@gmail.com>	2014-01-10 14:53:38
Raf Guns <rafguns@gmail.com>	2013-12-12 11:39:43
Raf Guns <rafguns@gmail.com>	2013-12-12 11:39:06

# Data set history

<https://github.com/datasets/glwd>

Commits on Aug 3, 2014



**Fix JSON in datapackage.json (missing ,).**

rgrp authored on 3 Aug 2014



3fc36c8



Commits on Mar 18, 2014



**updated datapackage.json with file sizes, and info**

jalbertbowden authored on 18 Mar 2014



dee5c14



**glwd global lakes and wetlands database gis (.shp) data, levels 1 and...** ...

jalbertbowden authored on 18 Mar 2014



071af08



**Update README.md**

jalbertbowden authored on 18 Mar 2014



d59250c



**Initial commit**

jalbertbowden authored on 18 Mar 2014

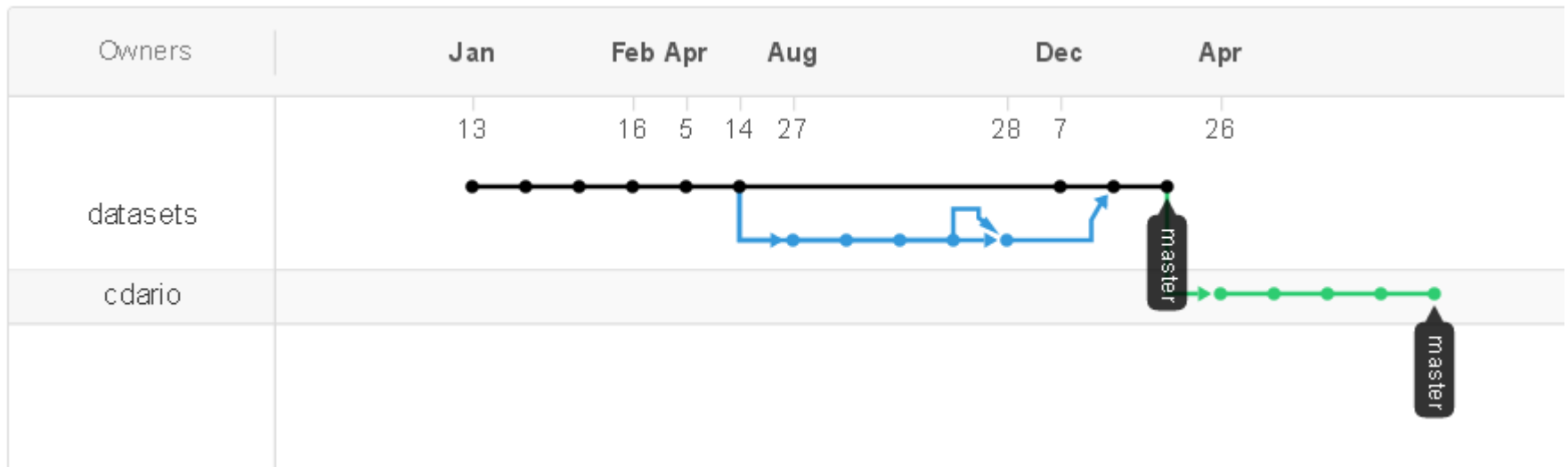


4e5a189



# Data set branches

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



# Data set 'diff'

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

350	350	JCI, Jemey (J.C.), Consumer Discretionary
351		-PNR, Pentair Ltd.
	351	+PNR, Pentair Ltd., Industrials
352	352	PBCT, People's United Bank, Financials
353	353	POM, Pepco Holdings Inc., Utilities
354	354	PEP, PepsiCo Inc., Consumer Staples
355	355	PKI, PerkinElmer, Health Care
356	356	PRGO, Perrigo, Health Care
357		-PETM, PETS MART Inc
	357	+PETM, PETS MART Inc, Consumer Discretionary
358	358	PFE, Pfizer Inc., Health Care



# 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”

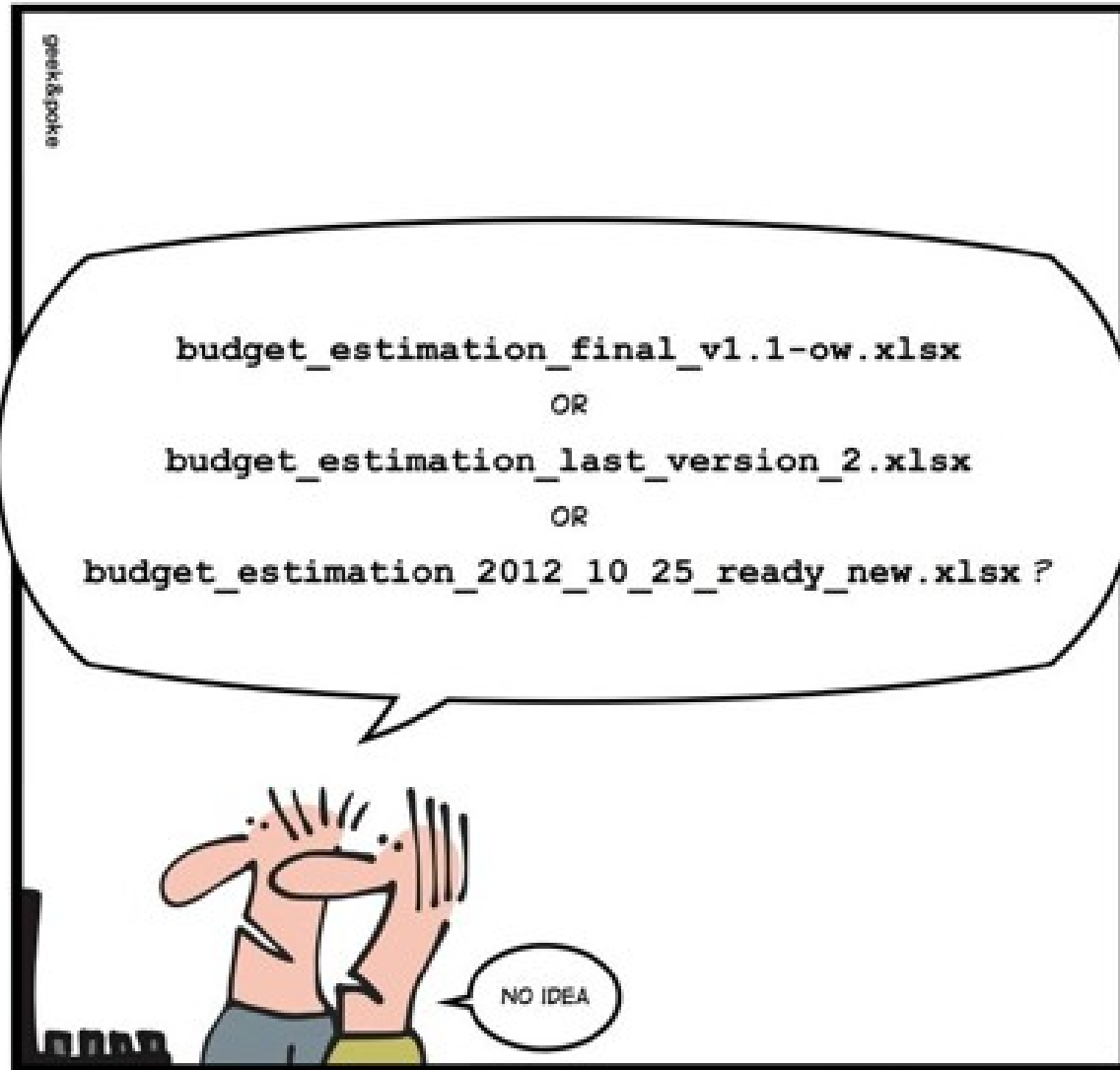
Two kinds of data:

- 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!



VERSION CONTROL