

# SINK

## SWISS ARMY KNIFE FOR DIRECTORY COMPARISON & SYNCHRONIZATION

Sébastien Pierre, Datalicious

**@Montréal Python 9, Sep. 2009**

[www.datalicious.ca](http://www.datalicious.ca) | [github.com/sebastien/sink](https://github.com/sebastien/sink)

**VERSION CONTROL**

**Is (TOO) GOOD !**

**BUT WE CAN'T USE  
IT FOR EVERYTHING**

**(LIKE MONITORING THE CHANGES AFTER `A make install`)**

**(LIKE WATCHING FOR LOCAL CHANGES IN A DIRECTORY)**

**(LIKE SAVING ONLY WHAT HAS CHANGED)**

# BUT WHY ?

**VCS**

**AREN'T FAST WITH BINARY FILES**

**MAKE A COPY OF YOUR DATA**

**ARE DESIGNED FOR FINE-GRAIN CHANGE DETECTION**

**NEED A REPOSITORY**

# **INTRODUCING**

## **SINK DIFF**

# GETTING PREPARED

```
$ git clone git://github.com/sebastien/sink.git sink-r1  
$ git clone git://github.com/sebastien/sink.git sink-r2  
$ git clone git://github.com/sebastien/sink.git sink-r3
```

We clone the same repository  
to **three folders**

```
$ sink diff sink-r1 sink-r2 sink-r3
```

**No differences**

```
$ cd sink-r1  
$ git checkout -b older a17609eaab  
$ cd ../sink-r2  
$ git checkout -b old 9e0e6a476f
```

# GETTING PREPARED

```
$ git clone git://github.com/sebastien/sink.git sink-r1  
$ git clone git://github.com/sebastien/sink.git sink-r2  
$ git clone git://github.com/sebastien/sink.git sink-r3
```

```
$ sink diff sink-r1 sink-r2 sink-r3  
No differences
```

We run **sink diff**  
and make sure they  
are the same

```
$ cd sink-r1  
$ git checkout -b older a17609eaab  
$ cd ../sink-r2  
$ git checkout -b old 9e0e6a476f
```

# GETTING PREPARED

```
$ git clone git://github.com/sebastien/sink.git sink-r1  
$ git clone git://github.com/sebastien/sink.git sink-r2  
$ git clone git://github.com/sebastien/sink.git sink-r3
```

```
$ sink diff sink-r1 sink-r2 sink-r3  
No differences
```

```
$ cd sink-r1  
$ git checkout -b older a17609eaab  
$ cd ../sink-r2  
$ git checkout -b old 9e0e6a476f
```

We update the repositories  
to **make sure they are**  
**at different revisions**

# SHOWING CHANGES

## **diff** sink-r1 sink-r2 sink-r3

```
00  ! [+] [+] .hgtags
01  [=] [-] [-] DESIGN
02  [=] [<] [<] Makefile
03  [=] [-] [-] NOTES
04  -!--!- [+] README
05  [=] [-] [-] ROADMAP
06  [=] [>] [<] TODO
07  -!- [+] [+] setup.py
08  -!- [+] -!- Documentation/DESIGN.txt
09  -!- [+] -!- Documentation/MANUAL.txt
10  [=] [=] [-] Resources/epydoc.css
11  -!- [+] [+] Scripts/sink
12  [=] [-] [-] Sources/sink/Sink.py
13  [=] [-] [-] Sources/sink/Tracking.py
14  -!- [+] [+] Sources/sink/linking.py
15  -!- [+] [+] Sources/sink/main.py
16  -!--!- [+] Sources/sink/snapshot.py
17  -!- [+] [+] Sources/sink/tracking.py
```

# SHOWING CHANGES

```
sink diff sink-r1 sink-r2 sink-r3
```

```
00 -!-[+][+] .hgtags
01 [=][-][-] DESIGN
02 [=][<][<] Makefile
03 [=][-][-] NOTES
04 -!--!-[+] README
05 [=][-][-] ROADMAP
06 [=][>][<] TODO
07 -!-[+][+] setup.py
08 -!-[+]-!- Documentation/DESIGN.txt
09 -!-[+]-!- Documentation/MANUAL.txt
10 [=][=][-] Resources/epydoc.css
11 -!-[+][+] Scripts/sink
12 [=][-][-] Sources/sink/Sink.py
13 [=][-][-] Sources/sink/Tracking.py
14 -!-[+][+] Sources/sink/linking.py
15 -!-[+][+] Sources/sink/main.py
16 -!--!-[+] Sources/sink/snapshot.py
17 -!-[+][+] Sources/sink/tracking.py
```

This file **was not**  
in the original version  
but **was added**  
in the R2 and R3

-!- MISSING FROM ORIGIN  
[+] MODIFIED

# SHOWING CHANGES

## `sink diff sink-r1 sink-r2 sink-r3`

```
00 -!-[+][+] .hgtags
01 [=][-][-] DESIGN
02 [=][<][<] Makefile
03 [=][-][-] NOTES
04 -!--!-[+] README
05 [=][-][-] ROADMAP
06 [=][<][<] TODO
07 -!-[+][+] setup.py
08 -!-[+]-!- Documentation/DESIGN.txt
09 -!-[+]-!- Documentation/MANUAL.txt
10 [=][=][-] Resources/epydoc.css
11 -!-[+][+] Scripts/sink
12 [=][-][-] Sources/sink/Sink.py
13 [=][-][-] Sources/sink/Tracking.py
14 -!-[+][+] Sources/sink/linking.py
15 -!-[+][+] Sources/sink/main.py
16 -!--!-[+] Sources/sink/snapshot.py
17 -!-[+][+] Sources/sink/tracking.py
```

This file **existed**  
in the original version  
but **was removed**  
in the R2 and R3

[=] SAME AS ORIGIN  
[-] REMOVED

# SHOWING CHANGES

## `sink diff sink-r1 sink-r2 sink-r3`

```
00 -!-[+][+] .hgtags
01 [=][-][-] DESIGN
02 [=][<][<] Makefile
03 [=][-][-] NOTES
04 -!--!-[+] README
05 [=][-][-] ROADMAP
06 [=][<][<] TODO
07 -!-[+][+] setup.py
08 -!-[+]-!- Documentation/DESIGN.txt
09 -!-[+]-!- Documentation/MANUAL.txt
10 [=][=][-] Resources/epydoc.css
11 -!-[+][+] Scripts/sink
12 [=][-][-] Sources/sink/Sink.py
13 [=][-][-] Sources/sink/Tracking.py
14 -!-[+][+] Sources/sink/linking.py
15 -!-[+][+] Sources/sink/main.py
16 -!--!-[+] Sources/sink/snapshot.py
17 -!-[+][+] Sources/sink/tracking.py
```

This file **existed**  
in the original version  
but **was modified**  
in the R2 and R3

[<]

DIFFERENT BUT OLDER

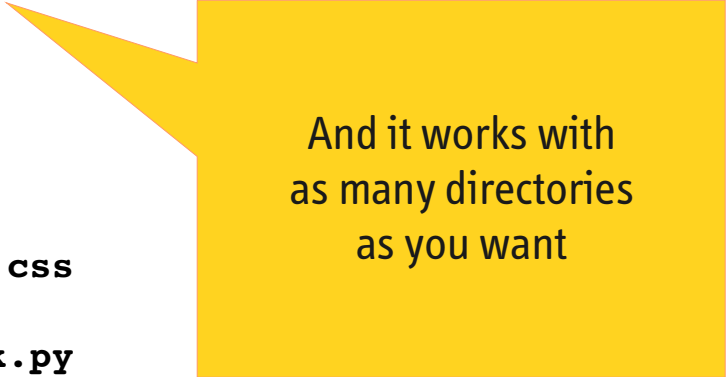
[>]

DIFFERENT AND NEWER

# SHOWING CHANGES

```
sink diff sink-r1 sink-r2 sink-r3 sink-r4 sink-r5 sink-r6 sink-r7  
sink-r8 sink-r9
```

```
00 -!-[+]-!--!-[+]-!--!--!-[+] .hgtags  
01 [=][>][<][<][<][<][<][<][>] Makefile  
02 [=][-][-][-][-][-][-][-][-] NOTES  
03 -!-[+][+][+][+][+][+][+][+][+] README  
04 [=][-][-][-][-][-][-][-][-] ROADMAP  
05 [=][<][<][<][<][<][<][<][<] TODO  
06 -!-[+][+][+][+][+][+][+][+][+] setup.py  
07 [=][-][-][-][-][-][-][-][-] Resources/epydoc.css  
08 -!-[+][+][+][+][+][+][+][+][+] Scripts/sink  
09 [=][-][-][-][-][-][-][-][-] Sources/sink/Sink.py  
10 [=][-][-][-][-][-][-][-][-] Sources/sink/Tracking.py  
11 -!-[+][+][+][+][+][+][+][+][+] Sources/sink/linking.py  
12 -!-[+][+][+][+][+][+][+][+][+] Sources/sink/main.py  
13 -!-[+][+][+][+][+][+][+][+][+] Sources/sink/snapshot.py  
14 -!-[+][+][+][+][+][+][+][+][+] Sources/sink/tracking.py
```



And it works with  
as many directories  
as you want





# UNIX FRIENDLY

```
sink diff +a +m sink-r1 sink-r2 | cut -d' ' -f3-
```

```
.hgtags  
README  
setup.py  
Scripts/sink  
Sources/sink/linking.py  
Sources/sink/main.py  
Sources/sink/snapshot.py  
Sources/sink/tracking.py
```

Use sink to extract a list of  
**changed or added files**

# UNIX FRIENDLY

```
sink diff +a +m sink-r1 sink-r2 | cut -d' ' -f3- \
```

```
| xargs tar cvfj mychanges.tar.bz2
```

```
.hgtags  
README  
setup.py  
Scripts/sink  
Sources/sink/linking.py  
Sources/sink/main.py  
Sources/sink/snapshot.py  
Sources/sink/tracking.py
```

Give them to tar and create  
a **tarball with the file-level  
diff**

# UNIX FRIENDLY

```
sink diff +a +m sink-r1 sink-r2 | cut -d' ' -f3- \
```

```
| xargs -I FILE cp sink-r2/FILE sink-r1/FILE
```

```
.hgtags  
README  
setup.py  
Scripts/sink  
Sources/sink/linking.py  
Sources/sink/main.py  
Sources/sink/snapshot.py  
Sources/sink/tracking.py
```

Or use it to (partly) **synchronize**  
**R1 with R2**



# STUFF YOU CAN DO

```
ssh myserver sink snap /etc > myserver-etc.snap  
sink diff myserver-etc.snap /etc
```

compare directories across  
different servers

```
{"rootLocation": "/etc", "accepts": [], "parent": null, "rejects": [], "attributes": {"Group": "0", "Creation": "1254342547", "Owner": "0", "Location": ".", "Modification": "1254342547", "Permissions": "16877", "Size": "12288"}, "type": "DirectoryNodeState", "children": [{"uid": "N365", "parent": ".", "rejects": [], "attributesSignature": false, "accepts": [], "attributes": {"Group": "0", "Creation": "1210072719", "Owner": "0", "Location": ".java", "Modification": "1210072719", "Permissions": "16877", "Size": "4096"}, "type": "DirectoryNodeState", "children": [{"uid": "N366", "parent": ".java", "rejects": [], "attributesSignature": false, "accepts": [], "attributes": {"Group": "0", "Creation": "1210072719", "Owner": "0", "Location": ".java/.systemPrefs", "Modification": "1210072719", "Permissions": "16877", "Size": "4096"}, "type": "DirectoryNodeState", "children": [{"uid": "N368", "parent": ".java/.systemPrefs", "rejects": [],
```

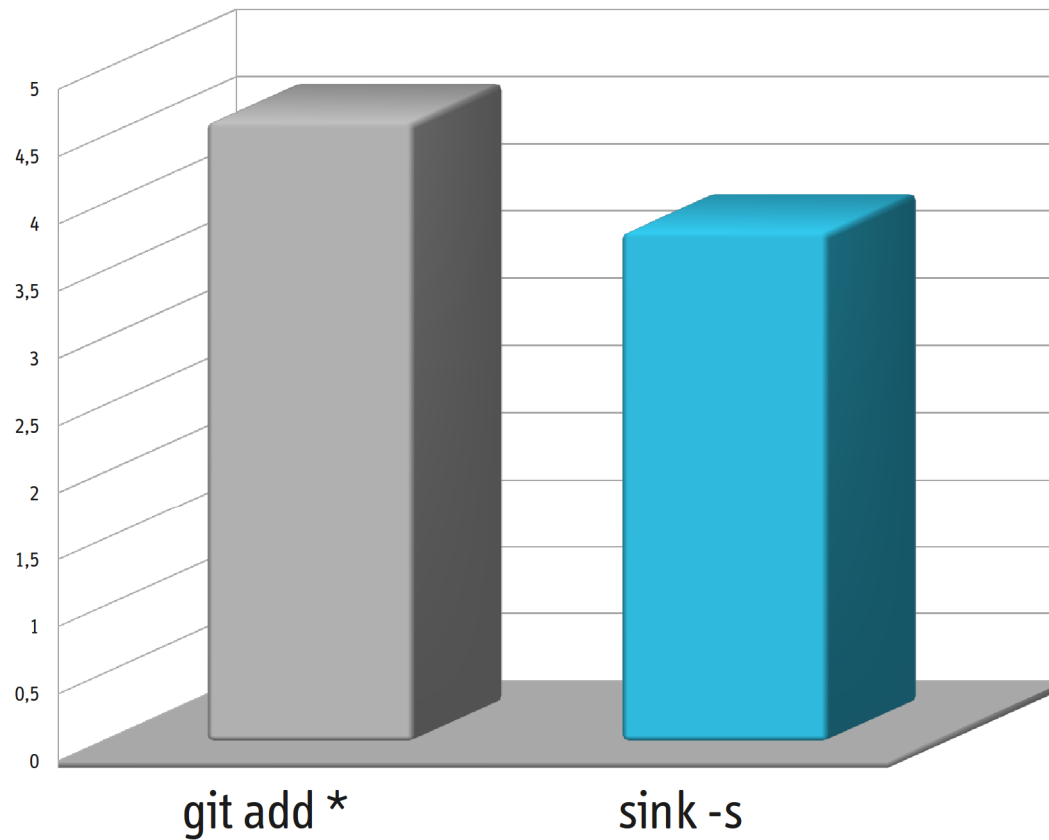
# STUFF YOU CAN DO

```
sink diff +a +m etc-20090928.snap /etc  
| xargs tar cvfj etc-updates-`date '+%Y%m%d'`.tar.bz2
```

Do an **incremental backup** of the changes you made to your configuration files

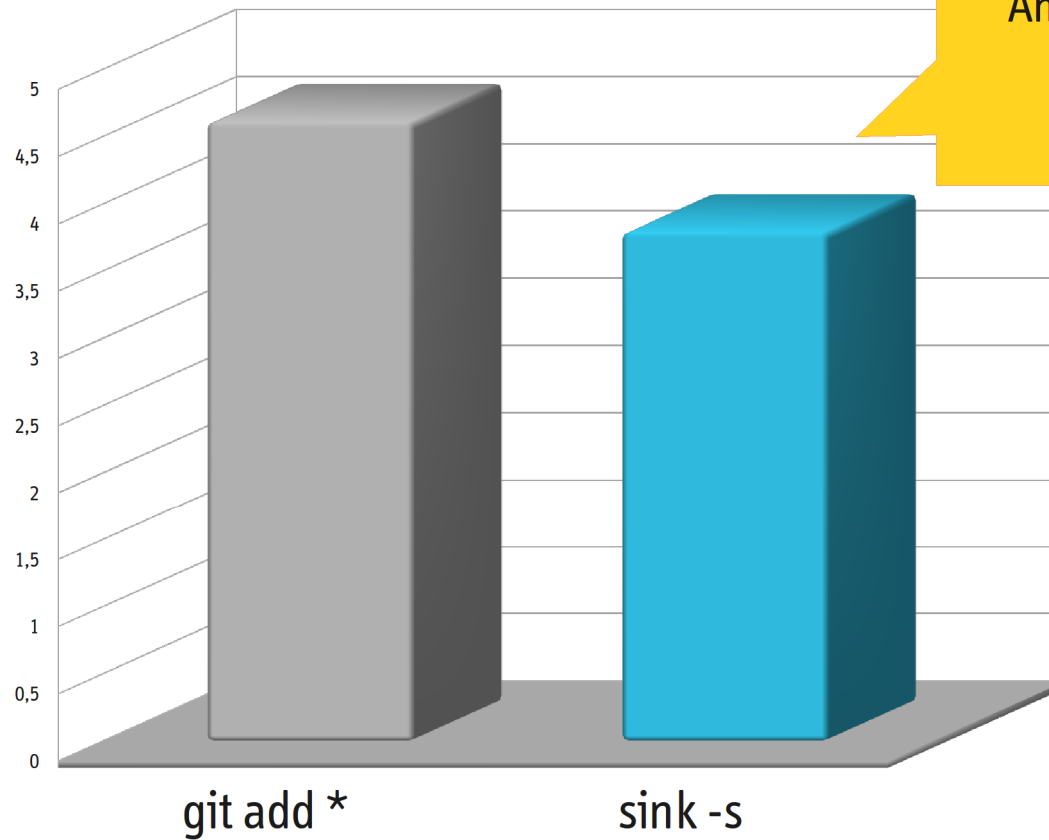
# SINK VS GIT (HAHA!)

**git add \* vs sink -s**  
**(lower is better, in seconds)**



# SINK VS GIT (HAHA!)

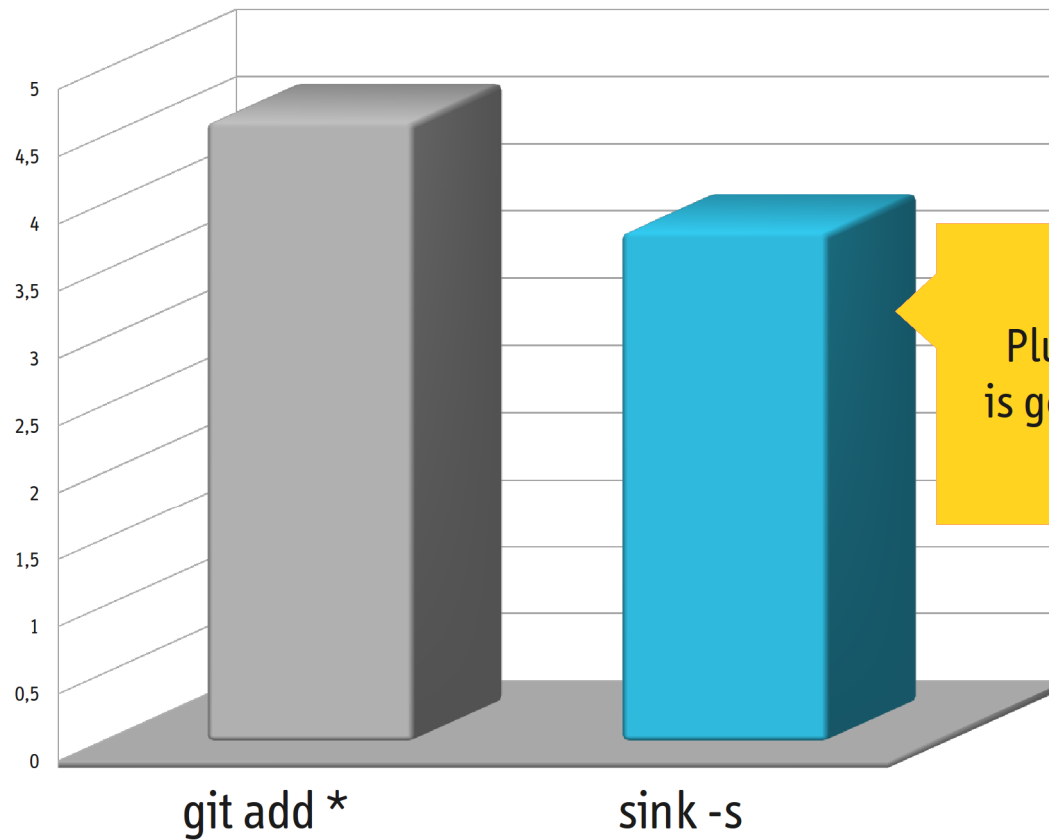
**git add \* vs sink -s**  
(lower is better, in seconds)



And it was on a large directory of 92Mb with 2955 files !

# SINK VS GIT (HAHA!)

**git add \* vs sink -s**  
**(lower is better, in seconds)**



Plus it is not that unfair as sink is getting the sha-1 for each file !

# SINK IN A NUTSHELL

## **Compare as many directories as you want**

Like diff -r on steroids !

## **Take snapshots of (remote) directories**

They track content and stat, plus it's small and JSON-encoded

## **UNIX friendly**

Go crazy with pipes and redirections to create mini-tools

## **It's got an API !**

It's so-so documented, but the code is rather clean !

**THE END**

**THANK YOU !**

**[WWW.GITHUB.COM/SEBASTIEN/SINK](http://WWW.GITHUB.COM/SEBASTIEN/SINK)**

**[WWW.DATALICIOUS.CA](http://WWW.DATALICIOUS.CA)  
[sebastien@datalicious.ca](mailto:sebastien@datalicious.ca)**