



R FOR SYSTEM ADMINISTRATION: AN EXAMPLE

LIGHTNING TALK @ *Open Source Open Mic*

Dirk Eddebuettel

August 3, 2017

The screenshot shows a web browser window with the URL `dirkeddelbuettel.com/cranberries/`. The page title is "CRANberries" and it has navigation links for "CRAN:New", "CRAN:Updated", "CRAN:Removed", "Misc", and "About".

Package BIEN updated to version 1.2.2 with previous version 1.2.1 dated 2017-07-31
Title: Tools for Accessing the Botanical Information and Ecology Network Database
Description: Provides Tools for Accessing the Botanical Information and Ecology Network Database. The BIEN database contains cleaned and standardized botanical data including occurrence, trait, plot and taxonomic data (See <http://Bien.nceas.ucsb.edu/bien/> for more Information). This package provides functions that query the BIEN database by constructing and executing optimized SQL queries.
Author: Brian Maitner [aut, cre]
Maintainer: Brian Maitner <brmaitner@gmail.com>

Diff between BIEN versions 1.2.1 dated 2017-07-31 and 1.2.2 dated 2017-08-03

```
DESCRIPTION | 6 +++-
MDS | 24 ++++++-----
NEWS | 12 ++++++
R/BIEN.R | 23 ++++++-----
inst/doc/BIEN.html | 4 +-
inst/doc/BIEN_tutorial.R | 42 ++++++-----
inst/doc/BIEN_tutorial.Rmd | 18 ++++++-----
inst/doc/BIEN_tutorial.html | 16 ++++++-----
man/BIEN_ranges_box.Rd | 4 +-
man/BIEN_ranges_genus.Rd | 8 +++-
man/BIEN_ranges_intersect_species.Rd | 4 +-
man/BIEN_ranges_species.Rd | 7 +++-
vignettes/BIEN_tutorial.Rmd | 18 ++++++-----
13 files changed, 111 insertions(+), 75 deletions(-)
```

More information about BIEN at CRAN
[Permanent link](#)

New package BALCONY with initial version 0.1.5
Package: BALCONY
Type: Package
Title: Better ALignment CONsensus anaLYsis
Version: 0.1.5
Author: Michał Stolarczyk & Alicja Pluciennik
Maintainer: Michał Stolarczyk <stolarczyk.micha193@gmail.com>
Description: Facilitates the evolutionary analysis and structure conservation study of specified amino acids in proteins.
License: GPL
Encoding: UTF-8
LazyData: true
Depends: seqinr, Rpdb, scales, stats, base
RoxygenNote: 6.0.1
NeedsCompilation: no
Packaged: 2017-08-02 19:31:43 UTC; mstolarczyk
Repository: CRAN
Date/Publication: 2017-08-03 04:46:19 UTC
More information about BALCONY at CRAN
[Permanent link](#)

MOTIVATION II

The screenshot shows a Twitter profile for 'CRAN Package Updates' (@CRANberriesFeed). The profile picture is a circular image of red cranberries. The header is blue with the Twitter logo and navigation icons. The profile bio includes the location 'Chicago, IL', the website 'dirk.eddelbuettel.com/cranberries/in...', and the join date 'January 2011'. There are 23K tweets and 1,009 followers. The main content area shows a list of tweets, each starting with 'CRAN Package Updates @CRANberriesFeed' and followed by a tweet about a new CRAN package update, such as 'New CRAN package BALCONY with initial version 0.1.5' and 'New CRAN package snpReady with initial version 0.9.3'. Each tweet includes a small cranberry profile picture icon and icons for reply, retweet, like, and share.

CRAN Package Updates
@CRANberriesFeed
Chicago, IL
dirk.eddelbuettel.com/cranberries/in...
Joined January 2011
26 Followers you know
Who to follow: Inside-R Community, Hilary Parker, R Tips

Tweets 23K Followers 1,009 Following

Tweets Tweets & replies

- CRAN Package Updates @CRANberriesFeed · 1h
CRAN updates: cranlike [goo.gl/5g3z](#) #rstats
- CRAN Package Updates @CRANberriesFeed · 2h
CRAN updates: eggCounts [glmBfp.goo.gl/5g3z](#) #rstats
- CRAN Package Updates @CRANberriesFeed · 5h
CRAN updates: SKAT [goo.gl/5g3z](#) #rstats
- CRAN Package Updates @CRANberriesFeed · 7h
CRAN updates: BIEN EfficientMaxEigenpair [jmvcore.goo.gl/5g3z](#) #rstats
- CRAN Package Updates @CRANberriesFeed · 7h
New CRAN package BALCONY with initial version 0.1.5 [goo.gl/pgjT](#) #rstats
- CRAN Package Updates @CRANberriesFeed · 16h
CRAN updates: GGally NPflow [goo.gl/5g3z](#) #rstats
- CRAN Package Updates @CRANberriesFeed · 16h
New CRAN package snpReady with initial version 0.9.3 [goo.gl/pgjT](#) #rstats

“COST”

```
eddbud: ~/cranberries/sources
eddbud:~/cranberries/sources$ ls -l *tar.gz | wc -l
57098
eddbud:~/cranberries/sources$ du -csh .
36G      .
36G      total
eddbud:~/cranberries/sources$
```

Over 57k files across 11k distinct packages

“PROBLEM”

```
edd@bud: ~/cranberries/sources
edd@bud:~/cranberries/sources$ ls -l A3*
-rw-rw-r-- 1 edd edd 45252 Feb  7  2013 A3_0.9.1.tar.gz
-rw-rw-r-- 1 edd edd 45890 Mar 25  2013 A3_0.9.2.tar.gz
-rw-rw-r-- 1 edd edd 42810 Aug 16  2015 A3_1.0.0.tar.gz
edd@bud:~/cranberries/sources$
```

For each of the 11k package ‘sets’, find the not-newest ones.

(ONE OF MANY POSSIBLE) SOLUTION(S)

`data.table`

- Highly-efficient `data.frame` extension
- Fast (by reference) aggregation/add/delete/join/group/...
- Operates via SQL-alike `dt[i, j, by]`
- More at r-datatable.com

STEP 1

```
## load package, silently
suppressMessages(library("data.table"))

## source directory
dir <- "~/cranberries/sources/"

## read files, optionally recursively (not here)
files <- list.files(dir, pattern="*tar.gz",
                    full.names=TRUE)
```

OUTPUT FROM STEP 1

```
R> suppressMessages(library("data.table"))
R> dir <- "~/cranberries/sources"
R> files <- list.files(dir, pattern="*tar.gz", full.names=TRUE)
R> d <- data.table(name=files, file.info(files))
R> d
```

	name	size	isdir	mode	mtime
1:	/home/edd/cranberries/sources/A3_0.9.1.tar.gz	45252	FALSE	664 2013-02-07 04:21:57.729160	
2:	/home/edd/cranberries/sources/A3_0.9.2.tar.gz	45890	FALSE	664 2013-03-25 02:12:03.437618	
3:	/home/edd/cranberries/sources/A3_1.0.0.tar.gz	42810	FALSE	664 2015-08-16 16:13:09.811910	
4:	/home/edd/cranberries/sources/aaMI_1.0-1.tar.gz	3487	FALSE	644 2005-10-17 14:24:18.000000	

57108:	/home/edd/cranberries/sources/ztype_0.1.0.tar.gz	3792	FALSE	664 2016-12-22 18:31:08.054666	
57109:	/home/edd/cranberries/sources/zyp_0.10-1.tar.gz	7280	FALSE	664 2013-09-19 02:12:34.938652	
57110:	/home/edd/cranberries/sources/zyp_0.9-1.tar.gz	5801	FALSE	644 2009-01-09 02:02:28.000000	
57111:	/home/edd/cranberries/sources/zyp_0.9-3.tar.gz	6959	FALSE	664 2013-08-23 18:12:29.087391	

	ctime	atime	uid	gid	uname	grname
1:	2017-06-23 07:28:42.504387	2017-06-25 19:36:52.266695	1000	1000	edd	edd
2:	2017-06-23 07:28:42.504387	2017-06-25 19:36:52.266695	1000	1000	edd	edd
3:	2017-06-23 07:28:42.660387	2017-06-25 19:36:52.274695	1000	1000	edd	edd
4:	2017-06-23 07:54:58.099709	2017-06-25 19:57:58.533000	1000	1000	edd	edd

57108:	2017-06-23 08:46:25.754497	2017-06-25 20:33:16.887890	1000	1000	edd	edd
57109:	2017-06-23 08:46:25.990497	2017-06-25 20:33:16.895890	1000	1000	edd	edd
57110:	2017-06-23 08:46:25.990497	2017-06-25 20:33:16.895890	1000	1000	edd	edd
57111:	2017-06-23 08:46:25.990497	2017-06-25 20:33:16.895890	1000	1000	edd	edd

```
R>
```


STEP 2

```
d[, baseNM := basename(name)]  
d[, nameNE := gsub(".tar.gz$", "", baseNM)]  
d[, pkg := gsub("(.)_.*$", "\\1", nameNE)]  
d[, ver := gsub(".*_(.*)$", "\\1", nameNE)]  
d
```

OUTPUT FROM STEP 2

```
R> d[, baseNM := basename(name)]
R> d[, nameNE := gsub(".tar.gz$", "", baseNM)]
R> d[, pkg := gsub("(.)_.*$", "\\1", nameNE)]
R> d[, ver := gsub(".*_(.*)$", "\\1", nameNE)]
R> d
```

	name	size	isdir	mode		mtime
1:	/home/edd/cranberries/sources/A3_0.9.1.tar.gz	45252	FALSE	664	2013-02-07	04:21:57.729160
2:	/home/edd/cranberries/sources/A3_0.9.2.tar.gz	45890	FALSE	664	2013-03-25	02:12:03.437618

57110:	/home/edd/cranberries/sources/zyp_0.9-1.tar.gz	5801	FALSE	644	2009-01-09	02:02:28.000000
57111:	/home/edd/cranberries/sources/zyp_0.9-3.tar.gz	6959	FALSE	664	2013-08-23	18:12:29.087391

	ctime	atime	uid	gid	uname	grname	baseNM
1:	2017-06-23 07:28:42.504387	2017-06-25 19:36:52.266695	1000	1000	edd	edd	A3_0.9.1.tar.gz
2:	2017-06-23 07:28:42.504387	2017-06-25 19:36:52.266695	1000	1000	edd	edd	A3_0.9.2.tar.gz

57110:	2017-06-23 08:46:25.990497	2017-06-25 20:33:16.895890	1000	1000	edd	edd	zyp_0.9-1.tar.gz
57111:	2017-06-23 08:46:25.990497	2017-06-25 20:33:16.895890	1000	1000	edd	edd	zyp_0.9-3.tar.gz

	nameNE	pkg	ver
1:	A3_0.9.1	A3	0.9.1
2:	A3_0.9.2	A3	0.9.2

57110:	zyp_0.9-1	zyp	0.9-1
57111:	zyp_0.9-3	zyp	0.9-3

```
R>
```

STEP 3

```
## set keys on package and mod.time
setkeyv(d, c("pkg", "mtime"))

## compute boolean on 'is it latest'
d[, newest := (ver == last(ver)), by=pkg]

## view 'name' of those that are not latest, cap'ed at 5
d[ newest==FALSE, name][1:5]
```

OUTPUT FROM STEP 3

```
R> setkeyv(d, c("pkg", "mtime"))
R>
R> d[ , newest := (ver == last(ver)), by=pkg]
R>
R> d[ newest==FALSE, name][1:5]
[1] "/home/edd/cranberries/sources/A3_0.9.1.tar.gz"
[2] "/home/edd/cranberries/sources/A3_0.9.2.tar.gz"
[3] "/home/edd/cranberries/sources/ABCanalysis_1.0.tar.gz"
[4] "/home/edd/cranberries/sources/ABCanalysis_1.0.1.tar.gz"
[5] "/home/edd/cranberries/sources/ABCanalysis_1.0.2.tar.gz"
R>
```

and those are our files to `unlink()` or move or ...

R

- lets us access standard POSIX functions
- in a vectorised manner

Operating 'sys-admin' style on files and folders

- is just [Programming with Data](#)
- for which R is ideally suited

Email

dirk@eddelbuettel.com

Website

<http://dirk.eddelbuettel.com>

Twitter

@eddelbuettel

GitHub

eddelbuettel