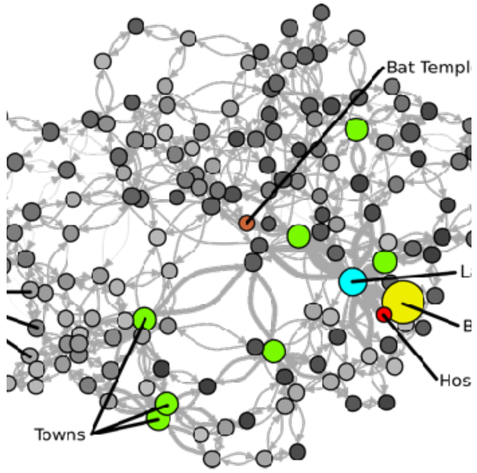
The Modern Data Package

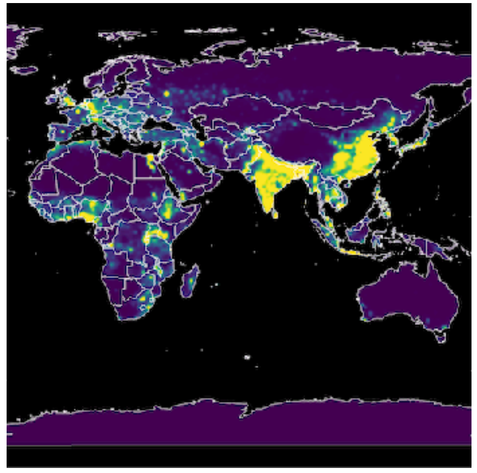
Noam Ross

rstats.nyc, 2017-04-22

noamross @

EcoHealth Alliance





EcoHealth Alliance

Local conservation. Global health.

@EcoHealthNYC ecohealthalliance.org

Building technical and community infrastructure for R to support open, reproducible science

@rOpenSci

ropensci.org

Best practices for open / scientific data

Machine- and human-readable formats

Future-proof, platform-agnostic, formats

Long-term archival repositories with fixed URIs

Standardized metadata

Documentation

Best practices for open / scientific data

~~Future-proof, platform-agnostic, formats~~

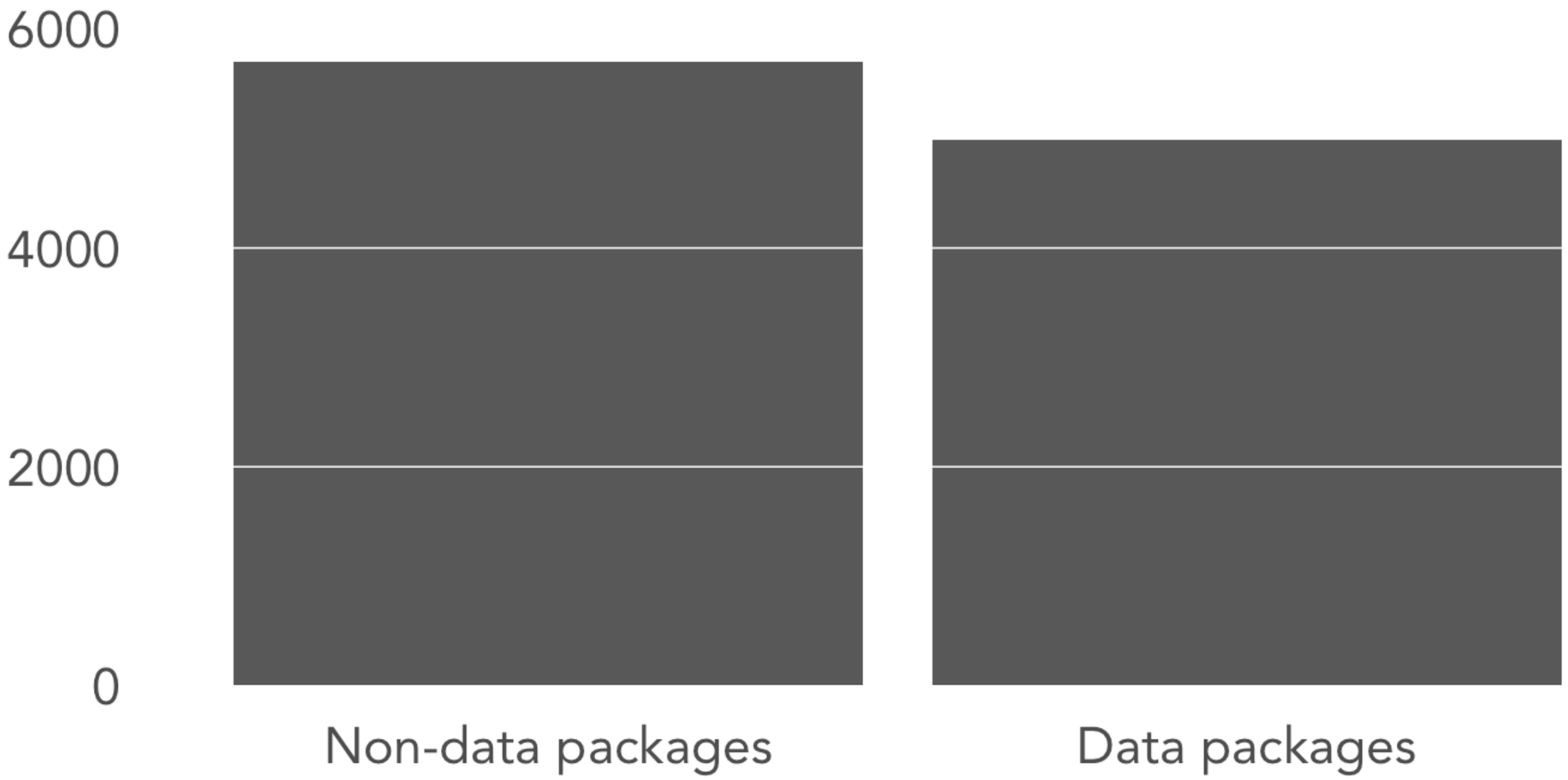
~~Machine- and human-readable formats~~

~~Long-term archival repositories with fixed URIs~~

Standardized metadata?

Documentation?

Data in CRAN Packages



github.com/ropenscilabs/data-packages

Advantages of a data package

Hijack package distribution and storage infrastructure

Take advantage of performance of non-archival formats

Facilitate anticipated workflows

Link data with software

Use software documentation systems

Data in packages	Data API wrappers	Data Sync
small data	unlimited size	local HD size
fixed data	live changes	regular updates
limited by hosting service, package structure	limited by transfer, server capacity, API endpoints	limited by local resources
teaching, demos	service interaction, targeted queries	bulk analysis

Data in packages	Data API wrappers	Data Sync
small data	unlimited size	local HD size
fixed data	live changes	regular updates
limited by hosting service, package structure	limited by transfer, server capacity, API endpoints	limited by local resources
teaching, demos	service interaction, targeted queries	bulk analysis

Data API Wrappers

PACKAGE	MAINTAINER	DESCRIPTION	DETAILS
AntWeb	Karthik Ram	Access data from the world's largest ant database. Maintained and developed by the California Academy of Science	CRAN
dbhydroR	Joseph Stachelek	Client for programmatic access to the South Florida Water Management District's DBHYDRO database	CRAN
europepmc	Najko Jahn	Access to the European PubMed Central database of papers and metadata	CRAN
nomisr	Evan Odell	Access Nomis UK Labour Market Data with R	CRAN
paleobioDB	Sara Varela	Access data from the Paleobiology Database, a warehouse of paleobiology database	CRAN
rbhl	Scott Chamberlain	Access full text and metadata on scanned and OCR text for biodiversity literature from Biodiversity Heritage Library	CRAN
rfishbase	Carl Boettiger	Access any fish data from Fishbase.org, including occurrence records, habitat data, and more	CRAN
rfisheries	Karthik Ram	Search and retrieve data from the OpenFisheries.org, currently providing access to global capture fishing landings from the Food and Agriculture Organization (FAO) of the United Nations	CRAN
rglobi	Jorrit Poelen	R library to access species interaction data of http://globalbioticinteractions.org	CRAN

ropensci.org/packages

Data in packages	Data API wrappers	Data Sync
small data	unlimited size	local HD size
fixed data	live changes	regular updates
limited by hosting service, package structure	limited by transfer, server capacity, API endpoints	limited by local resources
teaching, demos	service interaction, targeted queries	bulk analysis

Data Sync

Good for MB to GB scale data

Regular but not live updates, good for versioning

Speed and versatility local disks

Facilitate with package functions and docs

Versioning and Caching with datastorr

datastorr 0.0.3 Index

datastorr

Simple data versioning using GitHub to store data.

This package is designed to be used by other package authors, not directly by downstream end users.

End user interface

See here for the aim from the point of view for an end user.

They would install your package (which contains no data so is nice and light and can be uploaded to CRAN).

```
devtools::install_github("richfitz/datastorr.example")
```

The user can see what versions they have locally

```
datastorr.example::mydata_versions()
```

and can see what versions are present on github:

```
datastorr.example::mydata_versions(local=FALSE) # remote
```

To download the most recent dataset:

```
d <- datastorr.example::mydata()
```

Subsequent calls (even across R sessions) are cached so that the mydata() function is fast enough you can use it in place of the data.

To get a particular version:

```
d <- datastorr.example::mydata("0.0.1")
```

Downloads are cached across sessions using `rappdirs`.

Vignettes

- datastorr

Dependencies

- **Imports**: httr, rappdirs, storr
- **Suggests**: knitr, rmarkdown, testthat, whisker

Authors

@rgfitzjohn

github.com/ropenscilabs/datastorr

datastorr

Makes a data library with **rappdirs + storr**

Agnosticism to file type and method of loading

Autogeneration of code to include in package

Data versioning as GitHub releases

Potential for other back-ends (in data repositories)

Speed and Compression with fst

build passing | build passing | License AGPL v3 | CRAN 0.8.4 | codecov 46% | downloads 10K/month

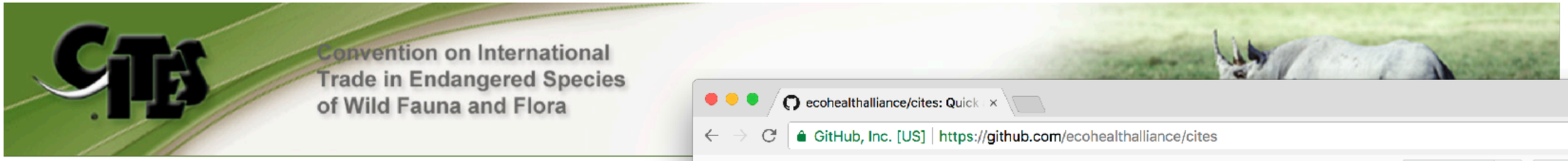
Overview

The *fst* package for R provides a fast, easy and flexible way to serialize data frames. With access speeds of multiple GB/s, *fst* is specifically designed to unlock the potential of high speed solid state disks that can be found in most modern computers. Data frames stored in the *fst* format have full random access, both in column and rows.

The figure below compares the read and write performance of the *fst* package to various alternatives.

Method	Format	Time (ms)	Size (MB)	Speed (MB/s)	N
readRDS	bin	1577	1000	633	112
saveRDS	bin	2042	1000	489	112
fread	csv	2925	1038	410	232
fwrite	csv	2790	1038	358	241
read_feather	bin	3950	813	253	112
write_feather	bin	1820	813	549	112
read_fst	**bin**	**457**	**303**	**2184**	**282**
write_fst	**bin**	**314**	**303**	**3180**	**291**

github.com/fstpackage/fst
github.com/krlmlr/fstplyr



CIT

Please enter your search below:

Year Range:
Exporting countries:
Importing countries:
Source:
Purpose:
Trade Terms:
Search by taxon:

YEAR RANGE :

Select start (From) and end (To) years for your query. It is
a year span not exceeding 5 years

From 2017 To 2017

ecohealthalliance/cites: Quick ×

Noam

GitHub, Inc. [US] | https://github.com/ecohealthalliance/cites

ecohealthalliance / cites

Watch 1 Star 1 Fork 0

Code Issues 2 Pull requests 0 Projects 0 Wiki Insights Settings

Quick access to all data in the CITES wildlife trade database https://trade.cites.org/ Edit

README.md

PASSED

cites

Authors: *Noam Ross*

The **cites** package provides a complete extract of the CITES wildlife trade database.

Installation

Install the **cites** package with this command:

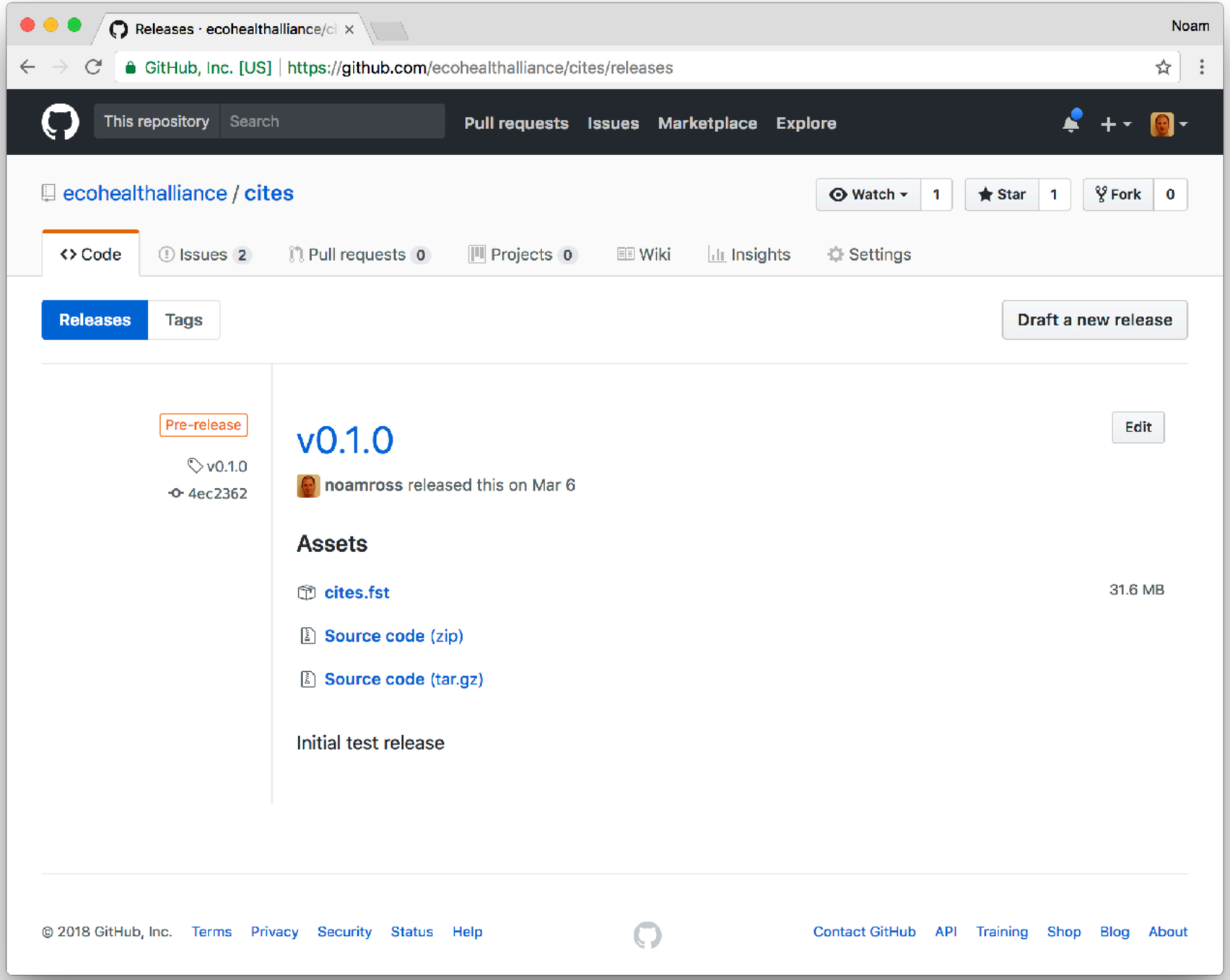
```
source("https://install-github.me/ecohealthalliance/cites")
```

Usage

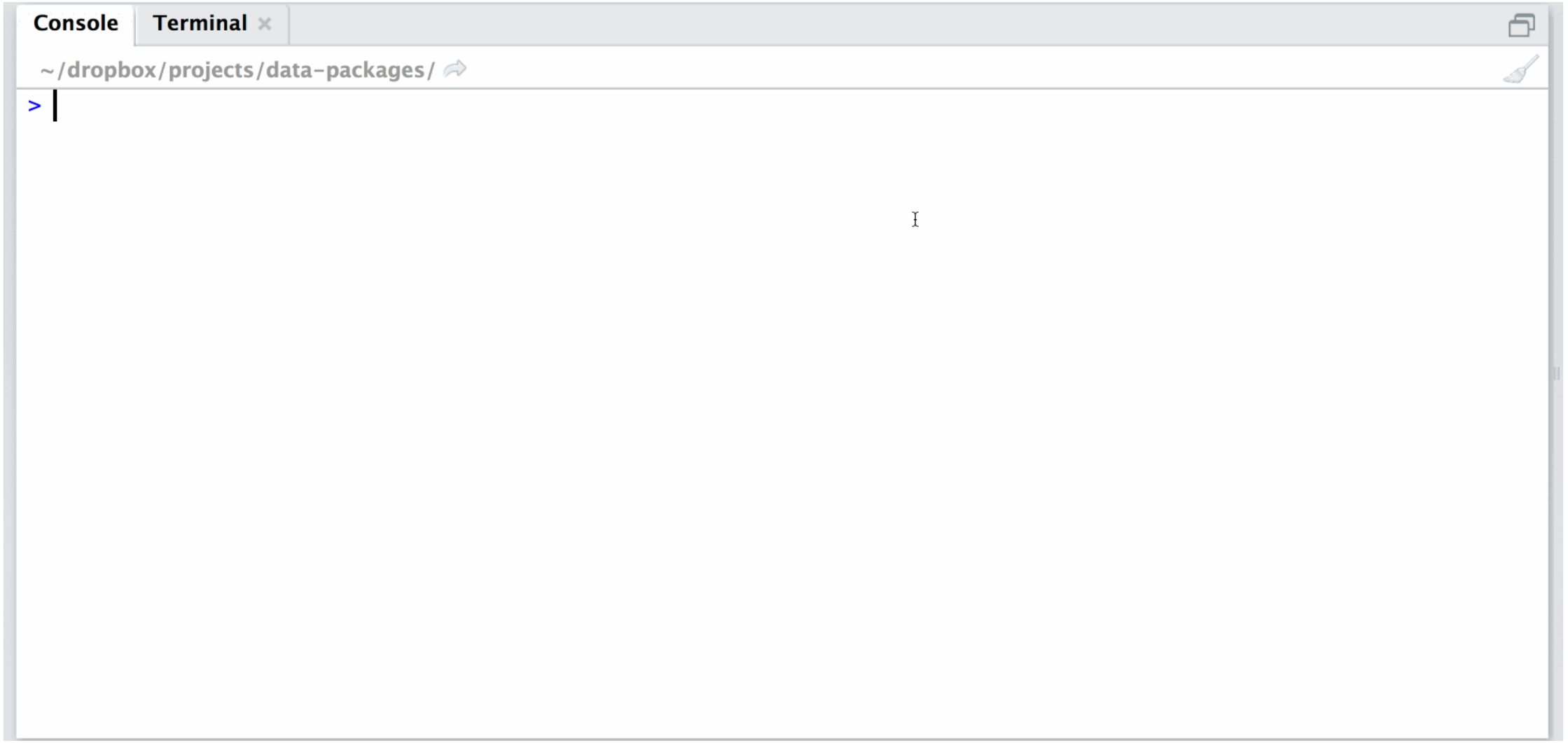
The main function in **cites** is `cites_data()`. This returns the main CITES database as a **dplyr** tibble.

cites makes use of **datastorr** to manage data download. The first time you run `cites_data()` the package will download the most recent version of the database (~32MB). Subsequent calls will load the database from storage on your computer.

trade.cites.org, github.com/ecohealthalliance/cites

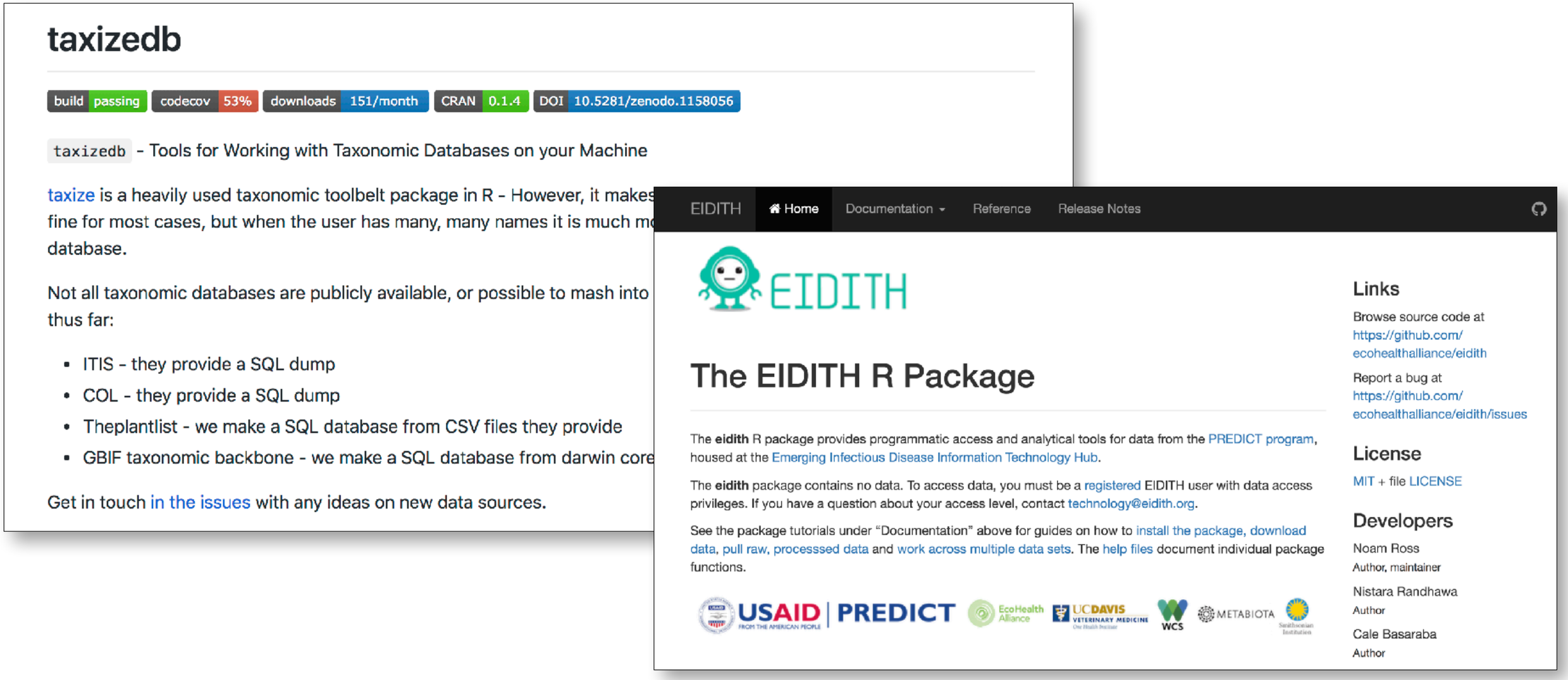


github.com/ecohealthalliance/cites



github.com/ecohealthalliance/cites

Syncing to Local Databases



taxizedb

build passing | codecov 53% | downloads 151/month | CRAN 0.1.4 | DOI 10.5281/zenodo.1158056

`taxizedb` - Tools for Working with Taxonomic Databases on your Machine

taxize is a heavily used taxonomic toolbelt package in R - However, it makes
fine for most cases, but when the user has many, many names it is much m
database.

Not all taxonomic databases are publicly available, or possible to mash into
thus far:

- ITIS - they provide a SQL dump
- COL - they provide a SQL dump
- Theplantlist - we make a SQL database from CSV files they provide
- GBIF taxonomic backbone - we make a SQL database from darwin core

Get in touch in the issues with any ideas on new data sources.

EIDITH | Home | Documentation | Reference | Release Notes

EIDITH

The EIDITH R Package

The **eidith** R package provides programmatic access and analytical tools for data from the PREDICT program, housed at the Emerging Infectious Disease Information Technology Hub.

The **eidith** package contains no data. To access data, you must be a registered EIDITH user with data access privileges. If you have a question about your access level, contact technology@eidith.org.

See the package tutorials under "Documentation" above for guides on how to install the package, download data, pull raw, processsed data and work across multiple data sets. The help files document individual package functions.

USAID | PREDICT · EcoHealth Alliance · UC DAVIS Veterinary Medicine · WCS · METABIOTA · Smithsonian Institution

Links

Browse source code at https://github.com/ecohealthalliance/eidith

Report a bug at https://github.com/ecohealthalliance/eidith/issues

License

MIT + file LICENSE

Developers

Noam Ross
Author, maintainer

Nistara Randhawa
Author

Cale Basaraba
Author

github.com/ropensci/taxizedb, github.com/ecohealthalliance/eidith

```
Console ~/dropbox-eha/projects/eidith/
> library(taxizedb)
> db_download_ncbi(overwrite = TRUE)
downloading...
unzipping...
loading 'names.dmp'...
|======================================================================================| 100%  145 MB
loading 'nodes.dmp'...
|======================================================================================| 100%  112 MB
building hierarchy table...
building SQLite database...
cleaning up...
[1] "/Users/noamross/Library/Caches/R/taxizedb/NCBI.sql"
>
```

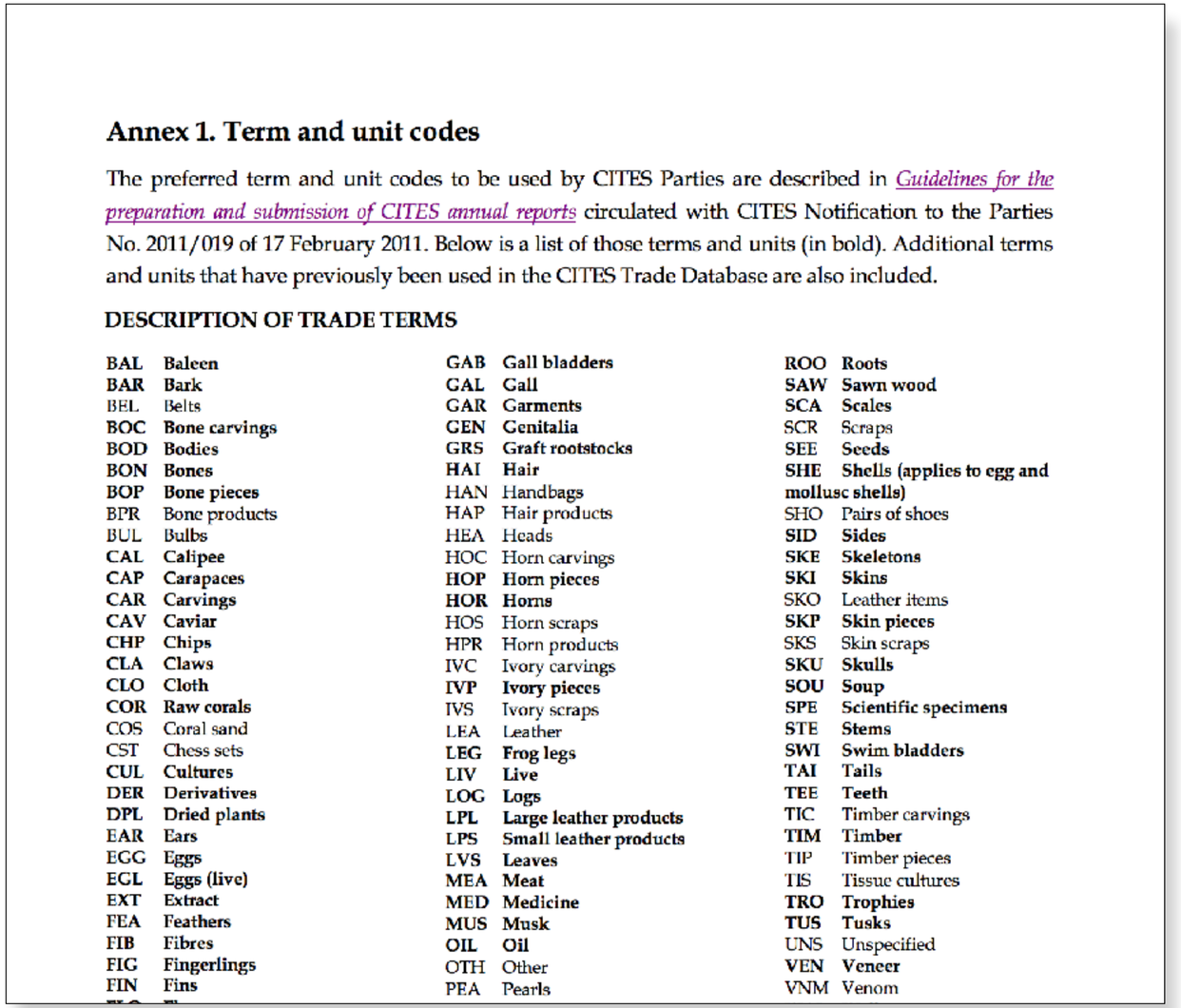
Console ~/dropbox-eha/projects/eidith/

```
> system.time(
+   taxize::children(3701, db='ncbi')
+ )
No ENTREZ API key provided
See https://ncbiinsights.ncbi.nlm.nih.gov/2017/11/02/new-api-keys-for-the-e-utilities/
No ENTREZ API key provided
See https://ncbiinsights.ncbi.nlm.nih.gov/2017/11/02/new-api-keys-for-the-e-utilities/
   user  system elapsed
  0.081   0.006   1.985
> system.time(
+   taxizedb::children(3701, db='ncbi')
+ )
   user  system elapsed
  0.124   0.021   0.144
>
```

Console ~/dropbox-eha/projects/eidith/

```
> library(eidith)
── EIDITH R Package ──────────────────────────────────────────────────────────
PREDICT-1 Table Status:
✔  Event Table                 Last Downloaded: 2018-04-10
✔  Animal Table                Last Downloaded: 2018-04-10
✔  Specimen Table              Last Downloaded: 2018-04-10
✔  Test Table                  Last Downloaded: 2018-04-10
✔  Virus Table                 Last Downloaded: 2018-04-10
✔  TestIDSpecimenID Table      Last Downloaded: 2018-04-10
PREDICT-2 Table Status:
✔  Event Table                          Last Downloaded: 2018-04-10
✔  Animal Table                         Last Downloaded: 2018-04-10
✔  Specimen Table                       Last Downloaded: 2018-04-10
✔  AnimalProduction Table               Last Downloaded: 2018-04-10
✔  CropProduction Table                 Last Downloaded: 2018-04-10
✔  Dwellings Table                      Last Downloaded: 2018-04-10
✔  ExtractiveIndustry Table             Last Downloaded: 2018-04-10
✔  MarketValueChain Table               Last Downloaded: 2018-04-10
✔  NaturalAreas Table                   Last Downloaded: 2018-04-10
✔  WildlifeRestaurant Table             Last Downloaded: 2018-04-10
✔  ZooSanctuary Table                   Last Downloaded: 2018-04-10
✔  Human Table                          Last Downloaded: 2018-04-10
✔  HumanCropProduction Table            Last Downloaded: 2018-04-10
✔  HumanAnimalProduction Table          Last Downloaded: 2018-04-10
✔  HumanExtractiveIndustry Table        Last Downloaded: 2018-04-10
✔  HumanHospitalWorker Table            Last Downloaded: 2018-04-10
✔  HumanHunter Table                    Last Downloaded: 2018-04-10
```

Making Metadata Useful



Annex 1. Term and unit codes

The preferred term and unit codes to be used by CITES Parties are described in *Guidelines for the preparation and submission of CITES annual reports* circulated with CITES Notification to the Parties No. 2011/019 of 17 February 2011. Below is a list of those terms and units (in bold). Additional terms and units that have previously been used in the CITES Trade Database are also included.

DESCRIPTION OF TRADE TERMS

BAL	**Baleen**
BAR	**Bark**
BEL	Belts
BOC	**Bone carvings**
BOD	**Bodies**
BON	**Bones**
BOP	**Bone pieces**
BPR	Bone products
BUL	Bulbs
CAL	**Calipee**
CAP	**Carapaces**
CAR	**Carvings**
CAV	**Caviar**
CHP	**Chips**
CLA	**Claws**
CLO	**Cloth**
COR	**Raw corals**
COS	Coral sand
CST	Chess sets
CUL	**Cultures**
DER	**Derivatives**
DPL	**Dried plants**
EAR	**Ears**
EGG	Eggs
EGL	**Eggs (live)**
EXT	**Extract**
FEA	**Feathers**
FIB	**Fibres**
FIG	**Fingerlings**
FIN	**Fins**
GAB	**Gall bladders**
GAL	**Gall**
GAR	**Garments**
GEN	**Genitalia**
GRS	**Graft rootstocks**
HAI	**Hair**
HAN	Handbags
HAP	Hair products
HEA	Heads
HOC	Horn carvings
HOP	**Horn pieces**
HOR	**Horns**
HOS	Horn scraps
HPR	Horn products
IVC	Ivory carvings
IVP	**Ivory pieces**
IVS	Ivory scraps
LEA	Leather
LEG	**Frog legs**
LIV	**Live**
LOG	**Logs**
LPL	**Large leather products**
LPS	**Small leather products**
LVS	**Leaves**
MEA	**Meat**
MED	**Medicine**
MUS	**Musk**
OIL	**Oil**
OTH	Other
PEA	Pearls
ROO	**Roots**
SAW	**Sawn wood**
SCA	**Scales**
SCR	Scraps
SEE	**Seeds**
SHE	**Shells (applies to egg and mollusc shells)**
SHO	Pairs of shoes
SID	**Sides**
SKE	**Skeletons**
SKI	**Skins**
SKO	Leather items
SKP	**Skin pieces**
SKS	Skin scraps
SKU	**Skulls**
SOU	**Soup**
SPE	**Scientific specimens**
STE	**Stems**
SWI	**Swim bladders**
TAI	**Tails**
TEE	**Teeth**
TIC	Timber carvings
TIM	**Timber**
TIP	Timber pieces
TIS	Tissue cultures
TRO	**Trophies**
TUS	**Tusks**
UNS	Unspecified
VEN	**Veneer**
VNM	Venom

github.com/ecohealthalliance/cites

Annex 1. Term and unit codes

The preferred term and unit codes to be used by CITES Parties are described in *Guidelines for the preparation and submission of CITES annual reports* circulated with CITES Notification to the Parties No. 2011/019 of 17 February 2011. Below is a list of those terms and units (in bold). Additional terms and units that have previously been used in the CITES Trade Database are also included.

DESCRIPTION OF TRADE TERMS

BAL	**Baleen**	**GAB**	**Gall bladders**	**ROO**	**Roots**
BAR	**Bark**	**GAL**	**Gall**	**SAW**	**Sawn wood**
BEL	Belts	**GAR**	**Garments**	**SCA**	**Scales**
BOC	**Bone carvings**	**GEN**	**Genitalia**	SCR	Scraps
BOD	**Bodies**	**GRS**	**Graft rootstocks**	**SEE**	**Seeds**
BON	**Bones**	**HAI**	**Hair**	**SHE**	**Shells (applies to egg and mollusc shells)**
BOP	**Bone pieces**	HAN	Handbags	
BPR	Bone products	HAP	Hair products	SHO	Pairs of shoes
BUL	Bulbs	HEA	Heads	**SID**	**Sides**
CAL	**Calipee**	HOC	Horn carvings	**SKE**	**Skeletons**
CAP	**Carapaces**	**HOP**	**Horn pieces**	**SKI**	**Skins**
CAR	**Carvings**	**HOR**	**Horns**	SKO	Leather items
CAV	**Caviar**	HOS	Horn scraps	**SKP**	**Skin pieces**
CHP	**Chips**	HPR	Horn products	SKS	Skin scraps
CLA	**Claws**	IVC	Ivory carvings	**SKU**	**Skulls**
CLO	**Cloth**	**IVP**	**Ivory pieces**	**SOU**	**Soup**
COR	**Raw corals**	IVS	Ivory scraps	**SPE**	**Scientific specimens**
COS	Coral sand	LEA	Leather	**STE**	**Stems**
CST	Chess sets	**LEG**	**Frog legs**	**SWI**	**Swim bladders**
CUL	**Cultures**	**LIV**	**Live**	**TAI**	**Tails**
DER	**Derivatives**	**LOG**	**Logs**	**TEE**	**Teeth**
DPL	**Dried plants**	**LPL**	**Large leather products**	TIC	Timber carvings
EAR	Ears	**LPS**	**Small leather products**	**TIM**	**Timber**
EGG	Eggs	**LVS**	**Leaves**	TIP	Timber pieces
EGL	**Eggs (live)**	**MEA**	**Meat**	TIS	Tissue cultures
EXT	**Extract**	**MED**	**Medicine**	**TRO**	**Trophies**
FEA	**Feathers**	**MUS**	**Musk**	**TUS**	**Tusks**
FIB	**Fibres**	**OIL**	**Oil**	UNS	Unspecified
FIG	**Fingerlings**	OTH	Other	**VEN**	**Veneer**
FIN	**Fins**	PEA	Pearls	VNM	Venom
FLO	**Flowers**	PIE	Pieces	WAL	Wallets
FOO	**Feet**	PKY	Piano keys	WAT	Watchstraps
FPT	**Flower pots**	**PLA**	**Plates**	**WAX**	**Wax**
FRA	Spectacle frames	**PLY**	**Plywood**	**WHO**	**Whole**
FRN	Items of furniture	**POW**	**Powder**	WOO	Wood products
FRU	**Fruit**	QUI	Quills	

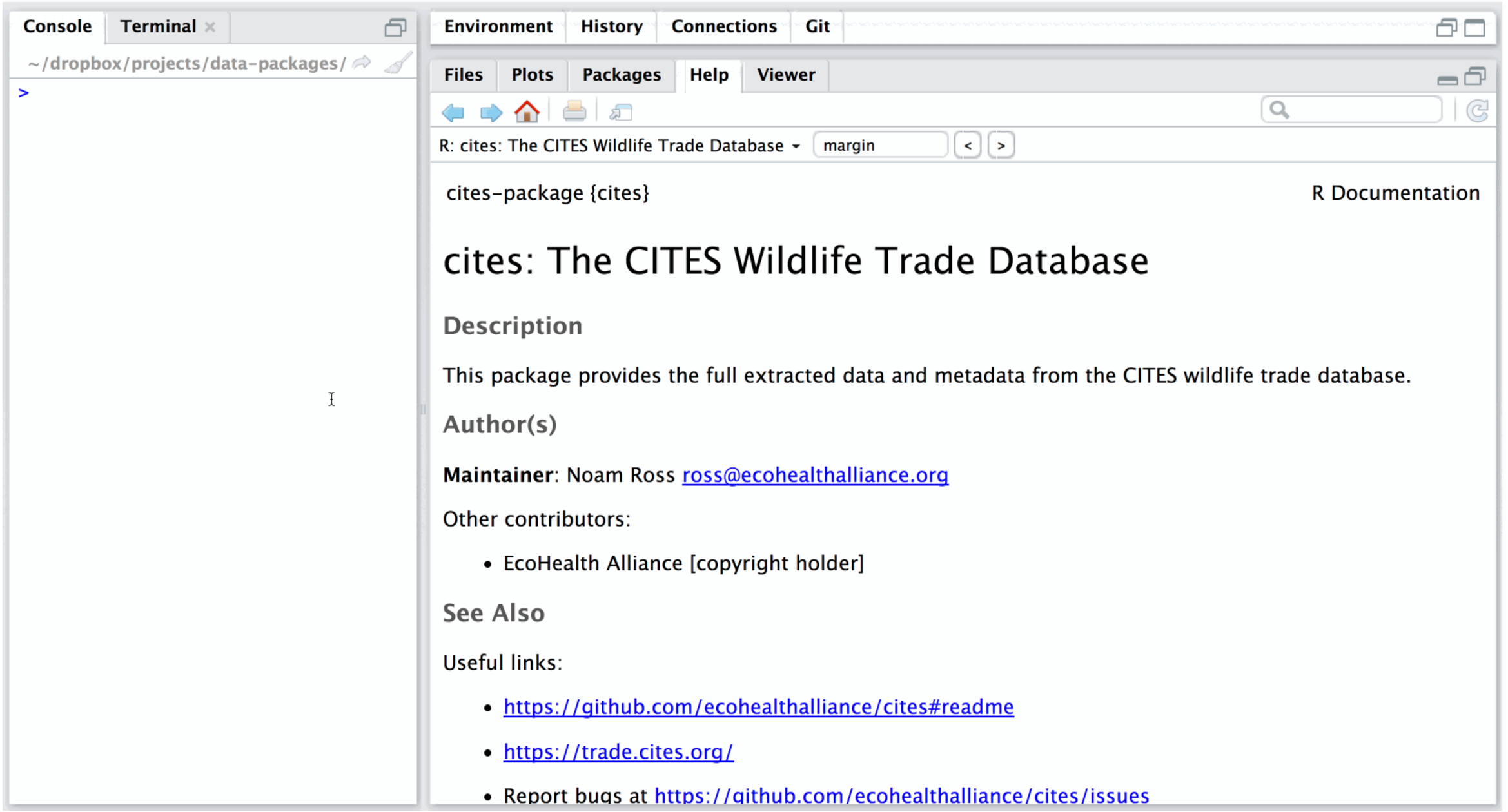
ropensci / tabulizer

Code | Issues 36 | Pull requests 2

Bindings for Tabula PDF Table Extractor Library

```
# A tibble: 165 x 3
   field code  description
   <chr> <chr> <chr>
 1 term  BAL   Baleen
 2 term  BAR   Bark
 3 term  BEL   Belts
 4 term  BOC   Bone carvings
 5 term  BOD   Bodies
 6 term  BON   Bones
 7 term  BOP   Bone pieces
 8 term  BPR   Bone products
 9 term  BUL   Bulbs
10 term  CAL   Calipee
# ... with 155 more rows
```

github.com/ecohealthalliance/cites, github.com/ropensci/tabulizer



github.com/ecohealthalliance/cites

Providing Analyst Breadcrumbs

Console ~/dropbox-eha/projects/eidith/

```
> library(eidith)
> ed2_human()

There are notes attached to this dataframe which may contain important information!
 Please look at the following column(s):
     livelihood_notes
     medical_history_notes
     movement_notes
# A tibble: 2,924 x 122
   event_name    biological_samp… participant_id date_of_intervi… begin_time_inte…
   <chr>         <chr>            <chr>          <chr>            <chr>
 1 BD-Dhaka Moh… no               BDAH0175       2017-12-13       9:30AM
 2 BD-Dhaka Moh… no               BDAH0176       2017-12-13       9:33AM
 3 BD-Dhaka Moh… no               BDAH0177       2017-12-13       10:45AM
 4 BD-Dhaka Moh… no               BDAH0178       2017-12-13       10:45AM
 5 BD-Dhaka Moh… no               BDAH0179       2017-12-13       12:15PM
 6 BD-Dhaka Moh… no               BDAH0180       2017-12-13       12:00PM
 7 BD-Dhaka Moh… no               BDAH0181       2017-12-18       10:05AM
 8 BD-Dhaka Moh… no               BDAH0182       2017-12-18       10:03AM
```

Useful patterns for data package design

Store, Query API or Sync? Consider data scale + turnover

Sync with versioned repositories

Use efficient/fast storage formats

Build rich documentation from structured metadata

Leave breadcrumbs to help analysts along the way

Minimize trade-offs with interoperability and archiving

Thanks!

noamross @ 

@EcoHealthNYC
ecohealthalliance.org

@rOpenSci
ropensci.org