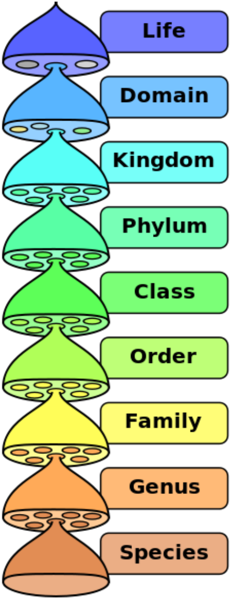
Taxa and metacoder: R packages for parsing, visualization, and manipulation of taxonomic data

Zachary Foster, Scott Chamberlain, Thomas Sharpton, and Niklaus Grunwald

The challenges of taxonomic data



- Taxonomic data is hierarchical
- Associated with tabular data
- Can be names, classifications, or IDs
- Many different taxonomic systems
- Many different data formats
- Hierarchical visualization is difficult

Sources of taxonomic data

DNA sequence databases

NCBI National Center for Biotechnology Information

unite

green genes
16S rRNA gene database and workbench compatible with ARB
greengenes.lbl.gov

Species occurrence databases

GBIF
Global Biodiversity Information Facility

iNaturalist.org

ATLAS OF LIVING AUSTRALIA
sharing biodiversity knowledge

Museum records

Smithsonian

北京自然博物馆
Beijing Museum of Natural History

NATURAL HISTORY MUSEUM

Sources of taxonomic data: DNA sequences

NCBI Genbank

AC073210.8 Homo sapiens BAC clone RP11-460N20 from 7, complete sequence

UNITE

SH099456.05FU_FJ357315_refs k__Fungi;p__Ascomycota;c__Dothideomycetes
;o__Pleosporales;f__Pleosporaceae;g__Embellisia;s__Embellisia_planifunda

RDP

S000448483 Sparassis crispa; MBUH-PIRJO&ILKKA94-1587/ss5
Lineage=Root;rootrank ;Fungi;domain;Basidiomycota;phylum;Agaricomycetes;
class;Polyporales;order ;Sparassidaceae;family;Sparassis;genus

SILVA

GCVF01000431.1.2369
Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales
;Alcanivoraceae;Alcanivorax;Thalassiosira rotula

Sources of taxonomic data: Occurrence records

Global Biodiversity Information Facility : Archea database

```r
readr::read_tsv("datasets/gbif_archea.csv")[4:8]
```

```
# A tibble: 19,013 x 5
   kingdom phylum        class        order             family
   <chr>   <chr>         <chr>        <chr>             <chr>
 1 Archaea Euryarchaeota Halobacteria Halobacteriales   Halobacteriaceae
 2 Archaea Euryarchaeota Thermococci  Thermococcales    Thermococcaceae
 3 Archaea Euryarchaeota Thermococci  Thermococcales    Thermococcaceae
 4 Archaea Crenarchaeota Thermoprotei Desulfurococcales Desulfurococcaceae
 5 Archaea Crenarchaeota Thermoprotei Desulfurococcales Pyrodictiaceae
 6 Archaea Crenarchaeota Thermoprotei Thermoproteales   Thermoproteaceae
 7 Archaea Euryarchaeota Thermococci  Thermococcales    Thermococcaceae
 8 Archaea Euryarchaeota Thermococci  Thermococcales    Thermococcaceae
 9 Archaea Euryarchaeota Halobacteria Halobacteriales   Halobacteriaceae
10 Archaea Euryarchaeota Halobacteria Halobacteriales   Halobacteriaceae
# ... with 19,003 more rows
```

Sources of taxonomic data: Museum records

Smithsonian Museum of Natural History: Mammal database

```r
readr::read_csv("datasets/SNMNH.csv")[9]
```

```
# A tibble: 5,000 x 1
   `Name Hierarchy`
   <chr>
 1 Abditomys latidens : Muridae : Rodentia : Mammalia : Chordata
 2 Abrawayaomys ruschii : Cricetidae : Rodentia : Mammalia : Chordata
 3 Abrawayaomys ruschii : Cricetidae : Rodentia : Mammalia : Chordata
 4 Abrocoma bennettii bennettii : Abrocomidae : Rodentia : Mammalia : Chordata
 5 Abrocoma bennettii bennettii : Abrocomidae : Rodentia : Mammalia : Chordata
 6 Abrocoma bennettii bennettii : Abrocomidae : Rodentia : Mammalia : Chordata
 7 Abrocoma bennettii bennettii : Abrocomidae : Rodentia : Mammalia : Chordata
 8 Abrocoma bennettii bennettii : Abrocomidae : Rodentia : Mammalia : Chordata
 9 Abrocoma bennettii bennettii : Abrocomidae : Rodentia : Mammalia : Chordata
10 Abrocoma bennettii bennettii : Abrocomidae : Rodentia : Mammalia : Chordata
# ... with 4,990 more rows
```

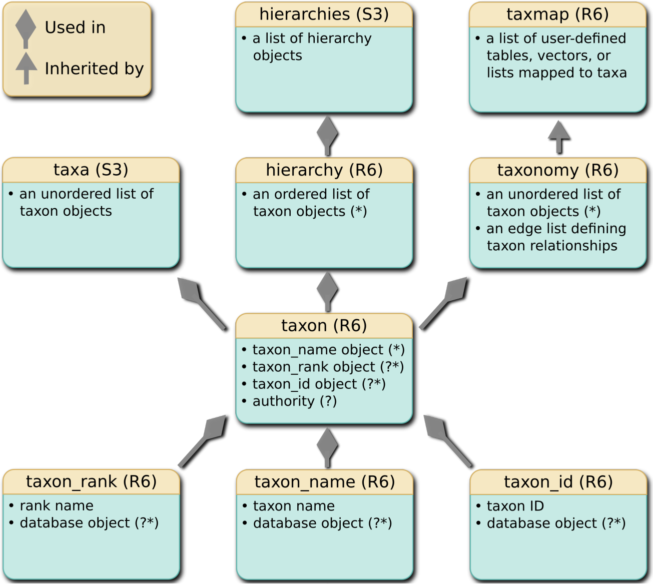
The `taxa` package

build passing | codecov 87% | repo status WIP | downloads 311/month | CRAN 0.2.1

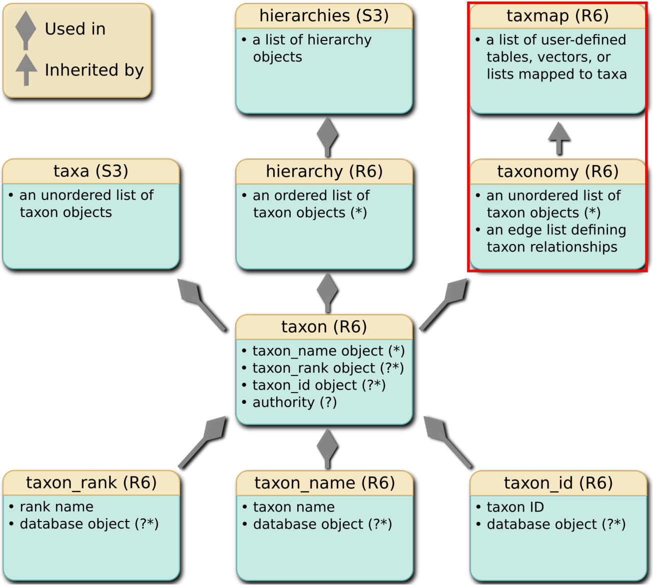
The `taxa` package is designed to be a soild foundation for using taxonomic data in R.

- Classes to hold taxa, taxonomies, and associated data
- Flexible parsers to convert raw data to these classes
- Dplyr-inspired functions to manipulate these classes
- Functions to get data assocaited with each taxon in a taxonomy

Classes defined by `taxa`: Relationships



Classes defined by `taxa`: Relationships



◆ Used in

↑ Inherited by

hierarchies (S3)

- a list of hierarchy objects

taxmap (R6)

- a list of user-defined tables, vectors, or lists mapped to taxa

taxa (S3)

- an unordered list of taxon objects

hierarchy (R6)

- an ordered list of taxon objects (*)

taxonomy (R6)

- an unordered list of taxon objects (*)
- an edge list defining taxon relationships

taxon (R6)

- taxon_name object (*)
- taxon_rank object (?*)
- taxon_id object (?*)
- authority (?)

taxon_rank (R6)

- rank name
- database object (?*)

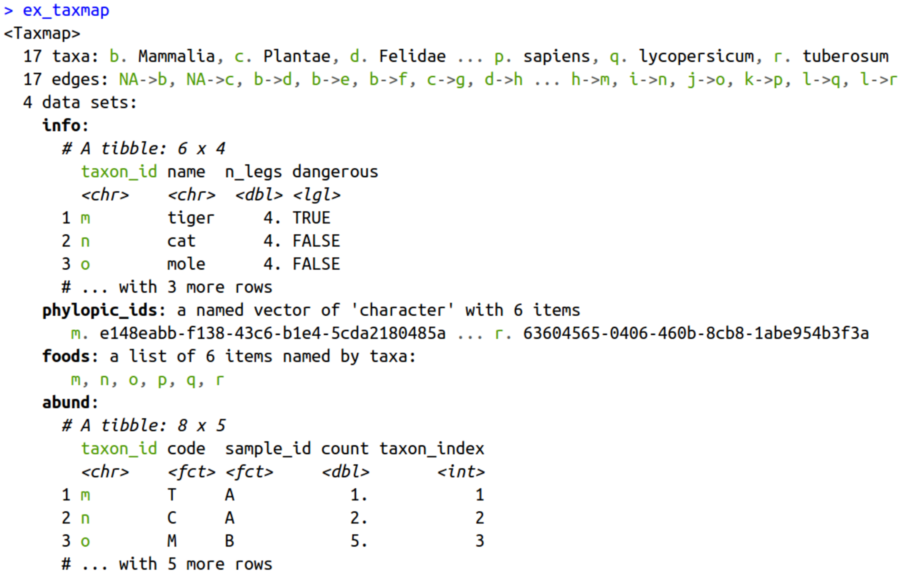
taxon_name (R6)

- taxon name
- database object (?*)

taxon_id (R6)

- taxon ID
- database object (?*)

Classes defined by taxa: The taxmap class



```
> ex_taxmap
<Taxmap>
  17 taxa: b. Mammalia, c. Plantae, d. Felidae ... p. sapiens, q. lycopersicum, r. tuberosum
  17 edges: NA->b, NA->c, b->d, b->e, b->f, c->g, d->h ... h->m, i->n, j->o, k->p, l->q, l->r
  4 data sets:
    info:
      # A tibble: 6 x 4
        taxon_id name  n_legs dangerous
        <chr>    <chr>  <dbl> <lgl>
      1 m        tiger     4. TRUE
      2 n        cat       4. FALSE
      3 o        mole      4. FALSE
      # ... with 3 more rows
    phylopic_ids: a named vector of 'character' with 6 items
       m. e148eabb-f138-43c6-b1e4-5cda2180485a ... r. 63604565-0406-460b-8cb8-1abe954b3f3a
    foods: a list of 6 items named by taxa:
       m, n, o, p, q, r
    abund:
      # A tibble: 8 x 5
        taxon_id code  sample_id count taxon_index
        <chr>    <fct> <fct>     <dbl>       <int>
      1 m        T     A            1.           1
      2 n        C     A            2.           2
      3 o        M     B            5.           3
      # ... with 5 more rows
```

Parsing

	Input data format		
Input type	Simple	Embedded	Raw string
	> print(data) [1] "**input_1**" "**input_2**" [3] "**input_3**"	> print(data) x **input** y 1 a **input_1** 100 2 b **input_2** 200 3 c **input_3** 300	> print(data) [1] ">id:a-tax:**input_1**" [2] ">id:b-tax:**input_2**" [3] ">id:c-tax:**input_3**"
Classification Primates;Hominidae;Homo;sapiens	> print(data) [1] "Primates;Hominidae;Hom... [2] "Primates;Haplorhini;Cr... > **parse_tax_data**(data, class_sep = ";")	> print(data) x class y 1 a Primates;Hominidae;... 100 2 b Primates;Haplorhini... 200 > **parse_tax_data**(data, class_cols = "class", class_sep = ";")	> print(data) [1] ">id:a-tax:Primates;Hom..." [2] ">id:b-tax:Primates;Hap..." > **extract_tax_data**(data, regex = ">id:(.+)-tax:(.+)", key = c("info", "class"), class_sep = ";")
Taxon ID 9606	> print(data) [1] "9606" "100937" ... > **lookup_tax_data**(data, type = "taxon_id")	> print(data) x id y 1 a 9606 100 2 b 100937 200 > **lookup_tax_data**(data, type = "taxon_id", column = "id")	> print(data) [1] ">id:a-tax:9606" [2] ">id:b-tax:100937" > **extract_tax_data**(data, regex = ">id:(.+)-tax:(.+)", key = c("info", "taxon_id"), database = "ncbi")
Taxon name Homo sapiens	> print(data) [1] "Homo sapiens" [2] "Primates" ... > **lookup_tax_data**(data, type = "taxon_name")	> print(data) x name y 1 a Homo sapiens 100 2 b Primates 200 > **lookup_tax_data**(data, type = "taxon_name", column = "name")	> print(data) [1] ">id:a-tax:Homo sapiens" [2] ">id:b-tax:Primates" > **extract_tax_data**(data, regex = ">id:(.+)-tax:(.+)", key = c("info","taxon_name"), database = "ncbi")
Sequence ID AC073210	> print(data) [1] "AC073210" "KC312885" ... > **lookup_tax_data**(data, type = "seq_id")	> print(data) x ncbi_id y 1 a AC073210 100 2 b KC312885 200 > **lookup_tax_data**(data, type = "seq_id", column = "ncbi_id")	> print(data) [1] ">id:a-tax:AC073210" [2] ">id:b-tax:KC312885" > **extract_tax_data**(data, regex = ">id:(.+)-tax:(.+)", key = c("info","seq_id"), database = "ncbi")

Parsing: vectors of classifications

```r
x <- c("Mammalia;Theria;Metatheria;Diprotodontia;Macropodiformes",
       "Mammalia;Theria;Eutheria;Primates;Haplorrhini;Simiiformes")

parse_tax_data(x, class_sep = ";")
```

```
## <Taxmap>
##   9 taxa: b. Mammalia, c. Theria ... j. Simiiformes
##   9 edges: NA->b, b->c, c->d, c->e, d->f, e->g, f->h, g->i, i->j
##   1 data sets:
##     tax_data: a named vector of 'character' with 2 items
##        h. Mammalia;Theria;[truncated] ... j. Mammalia;Theria;[truncated]
##   0 functions:
```

Parsing: vectors of names

```r
x <- c("Homo sapiens", "Macropus", "Chordata")

lookup_tax_data(x, type = "taxon_name", database = "ncbi")
```

```
## <Taxmap>
##   35 taxa: 131567. cellular organisms ... 9606. Homo sapiens
##   35 edges: NA->131567, 131567->2759 ... 207598->9605, 9605->9606
##   2 data sets:
##     tax_data:
##       # A tibble: 35 x 4
##         taxon_id ncbi_name          ncbi_rank    ncbi_id
##         <chr>    <chr>              <chr>        <chr>
##       1 131567   cellular organisms no rank      131567
##       2 2759     Eukaryota          superkingdom 2759
##       3 33154    Opisthokonta       no rank      33154
##       # ... with 32 more rows
##     query_data: a named vector of 'character' with 3 items
##        9606. Homo sapiens, 9312. Macropus, 7711. Chordata
##   0 functions:
```

Parsing: vectors of taxon IDs

```r
x <- c("9606", "207598", "7711") # NCBI taxon IDs

lookup_tax_data(x, type = "taxon_id", database = "ncbi")
```

```
## <Taxmap>
##   31 taxa: 131567. cellular organisms ... 9606. Homo sapiens
##   31 edges: NA->131567, 131567->2759 ... 207598->9605, 9605->9606
##   2 data sets:
##     tax_data:
##       # A tibble: 31 x 4
##         taxon_id ncbi_name          ncbi_rank    ncbi_id
##         <chr>    <chr>              <chr>        <chr>
##       1 131567   cellular organisms no rank      131567
##       2 2759     Eukaryota          superkingdom 2759
##       3 33154    Opisthokonta       no rank      33154
##       # ... with 28 more rows
##     query_data: a named vector of 'character' with 3 items
##        9606. 9606, 207598. 207598, 7711. 7711
##   0 functions:
```

Parsing: vectors of sequence IDs

```r
x <- c("AC073210", "MG014608", "AE006468") # NCBI sequence IDs

lookup_tax_data(x, type = "seq_id", database = "ncbi")
```

```
## <Taxmap>
##   46 taxa: 131567. cellular organisms ... 9606. Homo sapiens
##   46 edges: NA->131567, 131567->2759 ... 207598->9605, 9605->9606
##   2 data sets:
##     tax_data:
##       # A tibble: 46 x 4
##         taxon_id ncbi_name          ncbi_rank    ncbi_id
##         <chr>    <chr>              <chr>        <chr>
##       1 131567   cellular organisms no rank      131567
##       2 2759     Eukaryota          superkingdom 2759
##       3 33154    Opisthokonta       no rank      33154
##       # ... with 43 more rows
##     query_data: a named vector of 'character' with 3 items
##        9606. AC073210, 9316. MG014608, 99287. AE006468
##   0 functions:
```

Parsing: tables

Global Biodiversity Information Facility : Archea database

```
readr::read_tsv("datasets/gbif_archea.csv")[4:8]
```

```
# A tibble: 19,013 x 5
   kingdom phylum        class        order             family
   <chr>   <chr>         <chr>        <chr>             <chr>
 1 Archaea Euryarchaeota Halobacteria Halobacteriales   Halobacteriaceae
 2 Archaea Euryarchaeota Thermococci  Thermococcales    Thermococcaceae
 3 Archaea Euryarchaeota Thermococci  Thermococcales    Thermococcaceae
 4 Archaea Crenarchaeota Thermoprotei Desulfurococcales Desulfurococcaceae
 5 Archaea Crenarchaeota Thermoprotei Desulfurococcales Pyrodictiaceae
 6 Archaea Crenarchaeota Thermoprotei Thermoproteales   Thermoproteaceae
 7 Archaea Euryarchaeota Thermococci  Thermococcales    Thermococcaceae
 8 Archaea Euryarchaeota Thermococci  Thermococcales    Thermococcaceae
 9 Archaea Euryarchaeota Halobacteria Halobacteriales   Halobacteriaceae
10 Archaea Euryarchaeota Halobacteria Halobacteriales   Halobacteriaceae
# ... with 19,003 more rows
```

Parsing: tables

```r
x = readr::read_tsv("datasets/gbif_archea.csv")

parse_tax_data(x, class_cols = 4:8)
```

```
<Taxmap>
  95 taxa: ab. Archaea, ac. Euryarchaeota ... dr. Methermicoccaceae
  95 edges: NA->ab, ab->ac, ab->ad ... aw->dp, at->dq, ax->dr
  1 data sets:
    tax_data:
      # A tibble: 19,013 x 45
        taxon_id  gbifid datasetkey occurrenceid kingdom phylum class
        <chr>      <int> <chr>      <chr>        <chr>   <chr>  <chr>
      1 br         1.84e9 3c6e7390-~ D7C42A39-3A~ Archaea Eurya~ Halo~
      2 bs         1.83e9 863efcc4-~ <NA>         Archaea Eurya~ Ther~
      3 bs         1.83e9 863efcc4-~ <NA>         Archaea Eurya~ Ther~
      # ... with 1.901e+04 more rows, and 38 more variables:
      #   order <chr>, family <chr>, genus <chr>, species <chr>,
      #   infraspecificepithet <chr>, taxonrank <chr>,
      #   scientificname <chr>, countrycode <chr>, locality <chr>,
      #   publishingorgkey <chr>, ...
  0 functions:
```

Parsing: tables

Smithsonian Museum of Natural History: Mammal database

```
readr::read_csv("datasets/SNMNH.csv")[9]
```

```
# A tibble: 5,000 x 1
   `Name Hierarchy`
   <chr>
 1 Abditomys latidens : Muridae : Rodentia : Mammalia : Chordata
 2 Abrawayaomys ruschii : Cricetidae : Rodentia : Mammalia : Chordata
 3 Abrawayaomys ruschii : Cricetidae : Rodentia : Mammalia : Chordata
 4 Abrocoma bennettii bennettii : Abrocomidae : Rodentia : Mammalia : Chordata
 5 Abrocoma bennettii bennettii : Abrocomidae : Rodentia : Mammalia : Chordata
 6 Abrocoma bennettii bennettii : Abrocomidae : Rodentia : Mammalia : Chordata
 7 Abrocoma bennettii bennettii : Abrocomidae : Rodentia : Mammalia : Chordata
 8 Abrocoma bennettii bennettii : Abrocomidae : Rodentia : Mammalia : Chordata
 9 Abrocoma bennettii bennettii : Abrocomidae : Rodentia : Mammalia : Chordata
10 Abrocoma bennettii bennettii : Abrocomidae : Rodentia : Mammalia : Chordata
# ... with 4,990 more rows
```

Parsing: tables

```r
x = readr::read_csv("datasets/SNMNH.csv")

parse_tax_data(x, class_cols = "Name Hierarchy",
               class_sep = " : ", class_reversed = TRUE)
```

```
<Taxmap>
  141 taxa: ab. Chordata ... fl. Allactaga williamsi
  141 edges: NA->ab, ab->ac, ac->ad ... ay->fj, ay->fk, ay->fl
  1 data sets:
    tax_data:
      # A tibble: 5,000 x 53
        taxon_id Museum `Museum Abbreviation` `Catalog Number`
        <chr>    <chr>  <chr>                 <chr>
      1 az       <NA>   USNM                  357244
      2 ba       <NA>   USNM                  552416
      3 ba       <NA>   USNM                  <NA>
      # ... with 4,997 more rows, and 49 more variables: `Special
      #   Collections` <chr>, `Kind of Object` <chr>, `Specimen
      #   Count` <int>, `Current Identification` <chr>, `Other
      #   Identifications` <chr>, `Name Hierarchy` <chr>,
      #   Order <chr>, Family <chr>, `Type Status` <chr>, `Type
      #   Citations` <chr>, ...
  0 functions:
```

Parsing: complex strings (NCBI Genbank)

```
x = c("AC073210.8 Homo sapiens BAC clone RP11-460N20 from 7, complete sequence"
      "AE006468.2 Salmonella enterica subsp. enterica serovar Typhimurium",
      "MG014608.1 Macropus fuliginosus Csf1r gene, enhancer")

extract_tax_data(x, database = "ncbi", regex = "([A-Z0-9.]+) (.+)",
                 key = c(my_ncbi_id = "seq_id", my_desc = "info"))
```

```
<Taxmap>
  46 taxa: 131567. cellular organisms ... 9606. Homo sapiens
  46 edges: NA->131567, 131567->2759 ... 207598->9605, 9605->9606
  2 data sets:
    tax_data:
      # A tibble: 46 x 4
        taxon_id ncbi_name          ncbi_rank    ncbi_id
        <chr>    <chr>              <chr>        <chr>
      1 131567   cellular organisms no rank      131567
      2 2759     Eukaryota          superkingdom 2759
      3 33154    Opisthokonta       no rank      33154
      # ... with 43 more rows
    query_data:
      # A tibble: 3 x 4
        taxon_id my_ncbi_id my_desc              input
        <chr>    <chr>      <chr>                <chr>
      1 9606     AC073210.8 Homo sapiens BAC cl~ AC073210.8 Homo sap~
      2 99287    AE006468.2 Salmonella enterica~ AE006468.2 Salmonel~
```