## Sprint Review

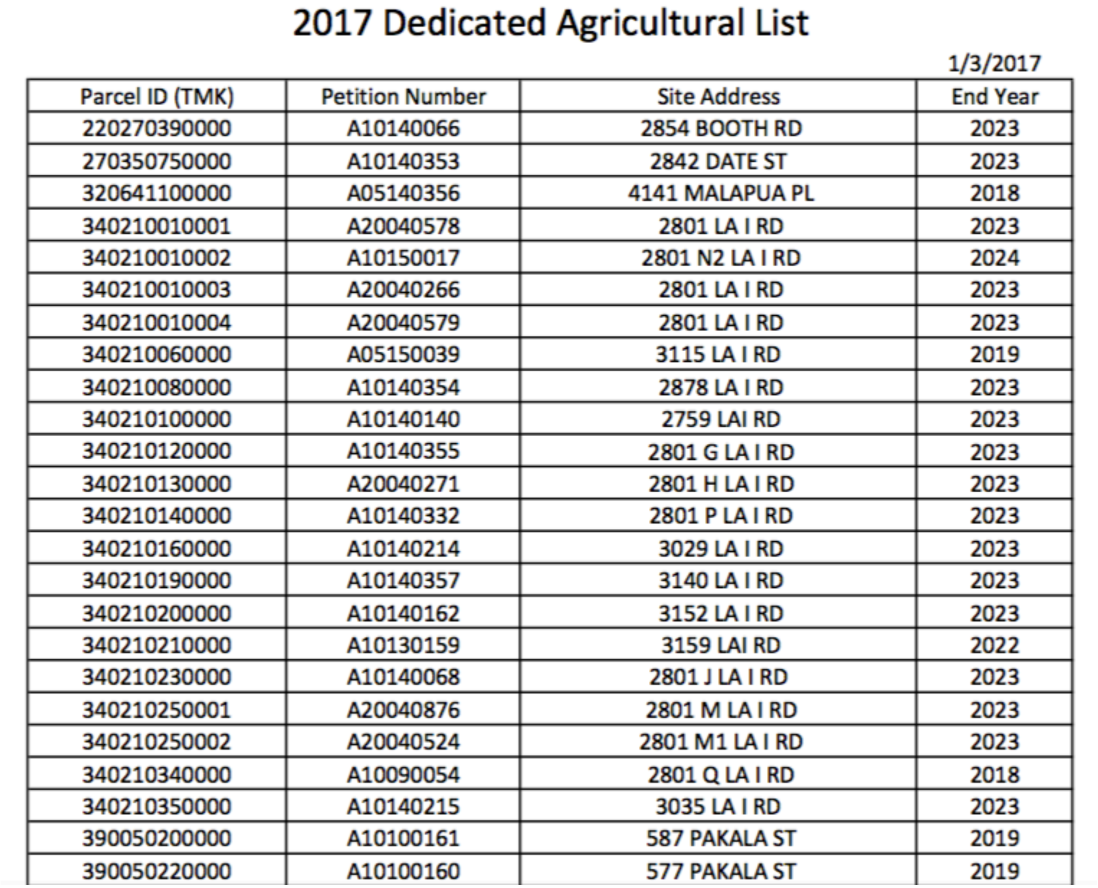
# Tables All The Way Down

☒ The image part with relationship ID rId3 was not found in the file.

# Tables in a PDF

- Often data is found in tables in a PDF
- That data wants to be free, but needs your help

www.shutterstock.com · 404305942



2017 Dedicated Agricultural List

1/3/2017

| Parcel ID (TMK) | Petition Number | Site Address | End Year |
|---|---|---|---|
| 220270390000 | A10140066 | 2854 BOOTH RD | 2023 |
| 270350750000 | A10140353 | 2842 DATE ST | 2023 |
| 320641100000 | A05140356 | 4141 MALAPUA PL | 2018 |
| 340210010001 | A20040578 | 2801 LA I RD | 2023 |
| 340210010002 | A10150017 | 2801 N2 LA I RD | 2024 |
| 340210010003 | A20040266 | 2801 LA I RD | 2023 |
| 340210010004 | A20040579 | 2801 LA I RD | 2023 |
| 340210060000 | A05150039 | 3115 LA I RD | 2019 |
| 340210080000 | A10140354 | 2878 LA I RD | 2023 |
| 340210100000 | A10140140 | 2759 LAI RD | 2023 |
| 340210120000 | A10140355 | 2801 G LA I RD | 2023 |
| 340210130000 | A20040271 | 2801 H LA I RD | 2023 |
| 340210140000 | A10140332 | 2801 P LA I RD | 2023 |
| 340210160000 | A10140214 | 3029 LA I RD | 2023 |
| 340210190000 | A10140357 | 3140 LA I RD | 2023 |
| 340210200000 | A10140162 | 3152 LA I RD | 2023 |
| 340210210000 | A10130159 | 3159 LAI RD | 2022 |
| 340210230000 | A10140068 | 2801 J LA I RD | 2023 |
| 340210250001 | A20040876 | 2801 M LA I RD | 2023 |
| 340210250002 | A20040524 | 2801 M1 LA I RD | 2023 |
| 340210340000 | A10090054 | 2801 Q LA I RD | 2018 |
| 340210350000 | A10140215 | 3035 LA I RD | 2023 |
| 390050200000 | A10100161 | 587 PAKALA ST | 2019 |
| 390050220000 | A10100160 | 577 PAKALA ST | 2019 |

The image part with relationship ID rId3 was not found in the file.

# Tables in a PDF

- There are various means to extract those tables:
- Applications, Web apps, R packages, etc.,
- They will often have different outputs (one vs multiple tables)

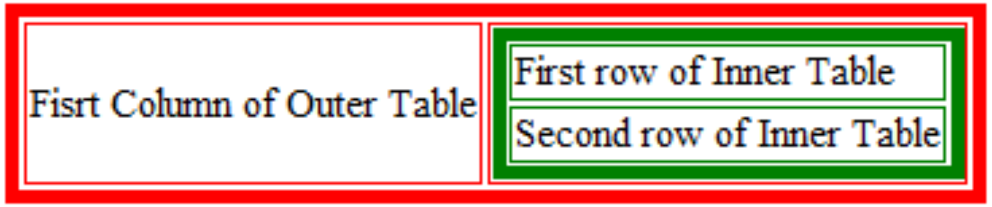
TABULA
liberate data tables trapped inside PDFs

PDFTables

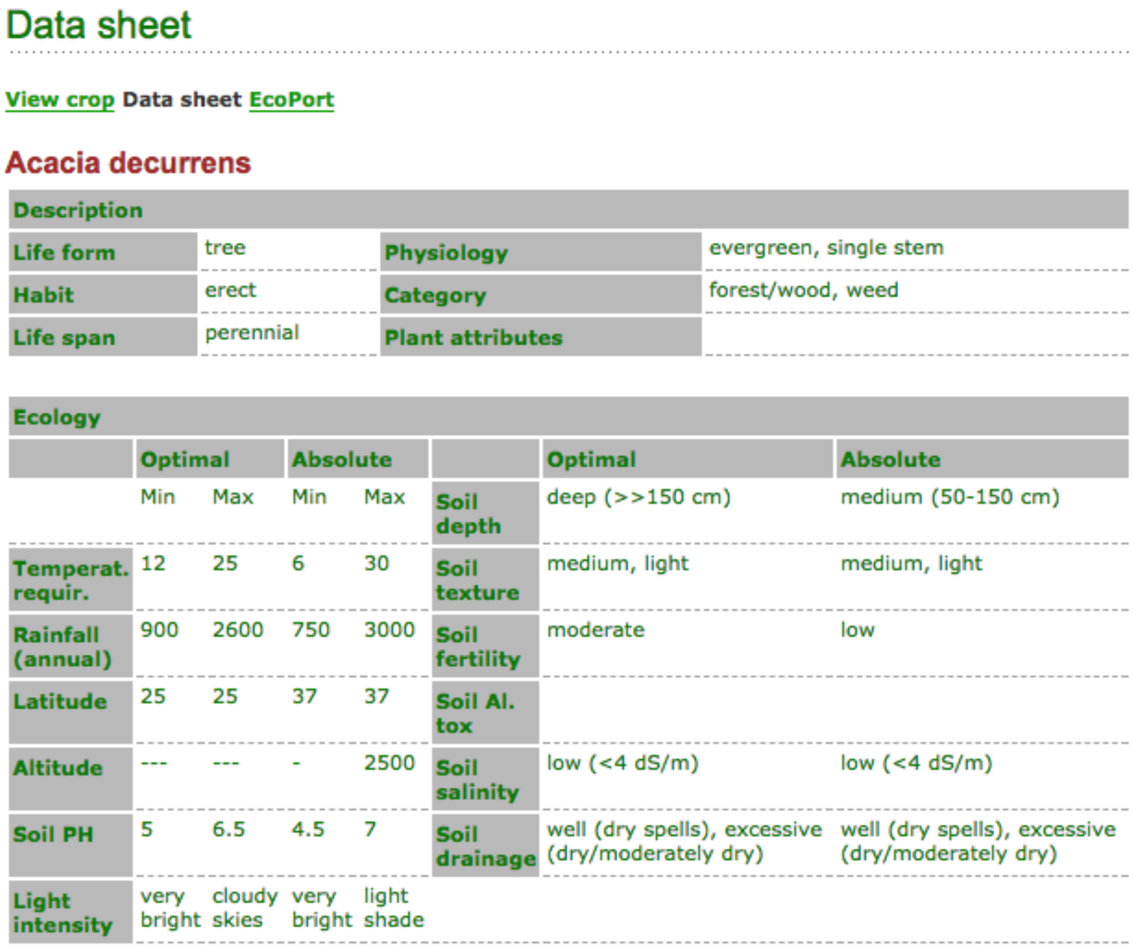
The image part with relationship ID rId3 was not found in the file.

# Tables in HTML

- **<tables> are also found in HTML code**
- Sometimes handily notated, often not
- You can even have tables inside tables



| | |
|---|---|
| Fisrt Column of Outer Table | First row of Inner Table<br>Second row of Inner Table |



## Data sheet

View crop Data sheet EcoPort

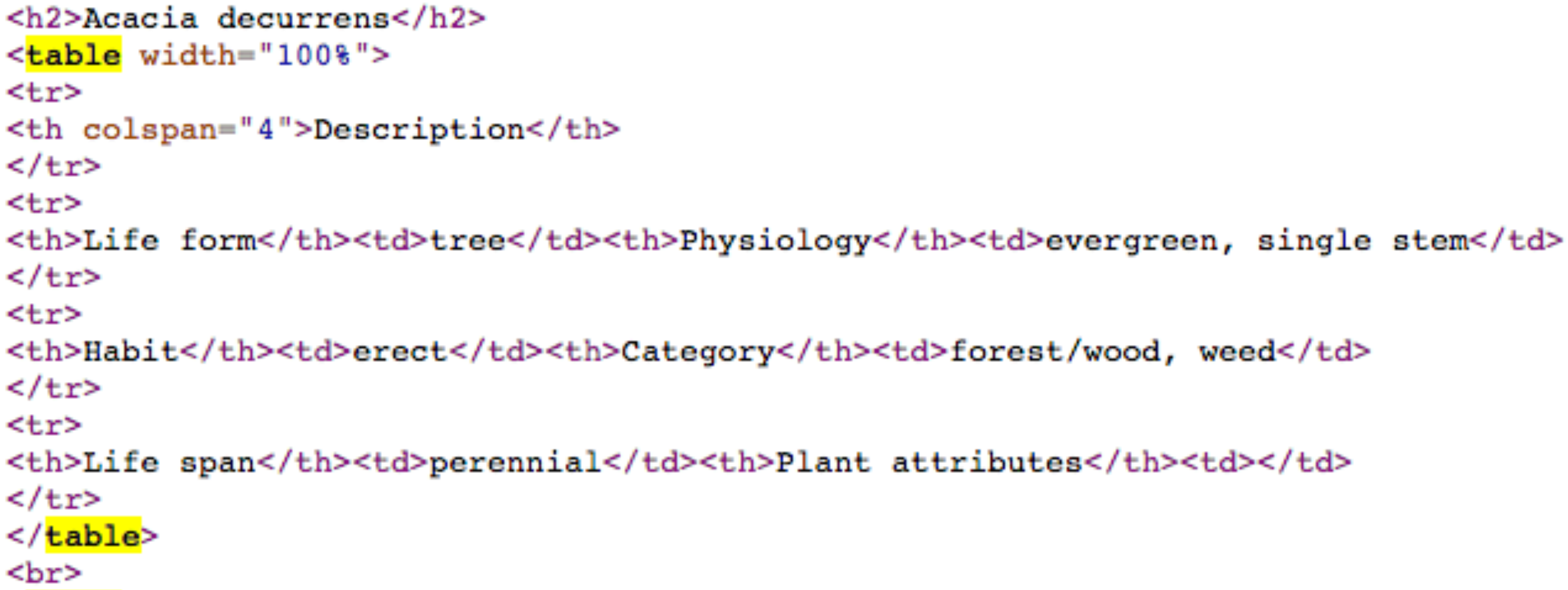
### Acacia decurrens

| Description | | | |
|---|---|---|---|
| Life form | tree | Physiology | evergreen, single stem |
| Habit | erect | Category | forest/wood, weed |
| Life span | perennial | Plant attributes | |

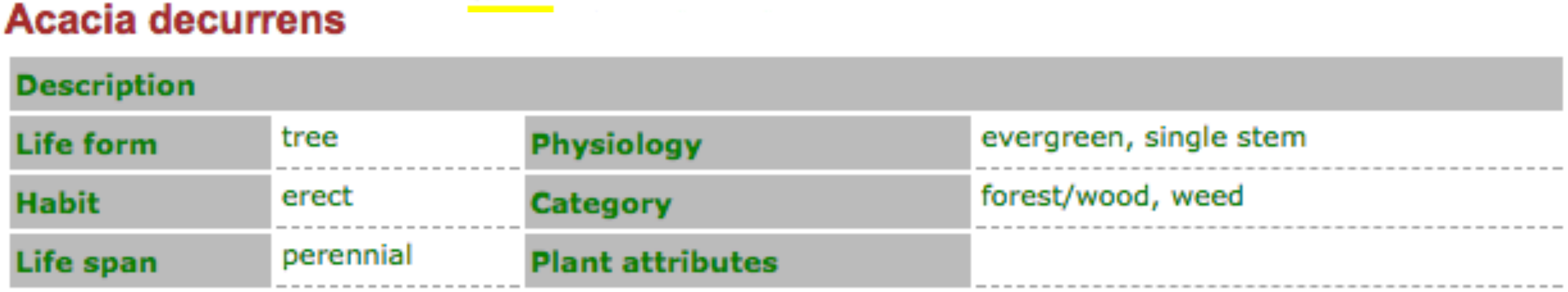
| Ecology | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Optimal | | Absolute | | | Optimal | Absolute |
| | Min | Max | Min | Max | Soil depth | deep (>>150 cm) | medium (50-150 cm) |
| Temperat. requir. | 12 | 25 | 6 | 30 | Soil texture | medium, light | medium, light |
| Rainfall (annual) | 900 | 2600 | 750 | 3000 | Soil fertility | moderate | low |
| Latitude | 25 | 25 | 37 | 37 | Soil Al. tox | | |
| Altitude | --- | --- | - | 2500 | Soil salinity | low (<4 dS/m) | low (<4 dS/m) |
| Soil PH | 5 | 6.5 | 4.5 | 7 | Soil drainage | well (dry spells), excessive (dry/moderately dry) | well (dry spells), excessive (dry/moderately dry) |
| Light intensity | very bright | cloudy skies | very bright | light shade | | | |

# Tables in HTML

- **<tables> are also found in HTML code**



```
<h2>Acacia decurrens</h2>
<table width="100%">
<tr>
<th colspan="4">Description</th>
</tr>
<tr>
<th>Life form</th><td>tree</td><th>Physiology</th><td>evergreen, single stem</td>
</tr>
<tr>
<th>Habit</th><td>erect</td><th>Category</th><td>forest/wood, weed</td>
</tr>
<tr>
<th>Life span</th><td>perennial</td><th>Plant attributes</th><td></td>
</tr>
</table>
<br>
```

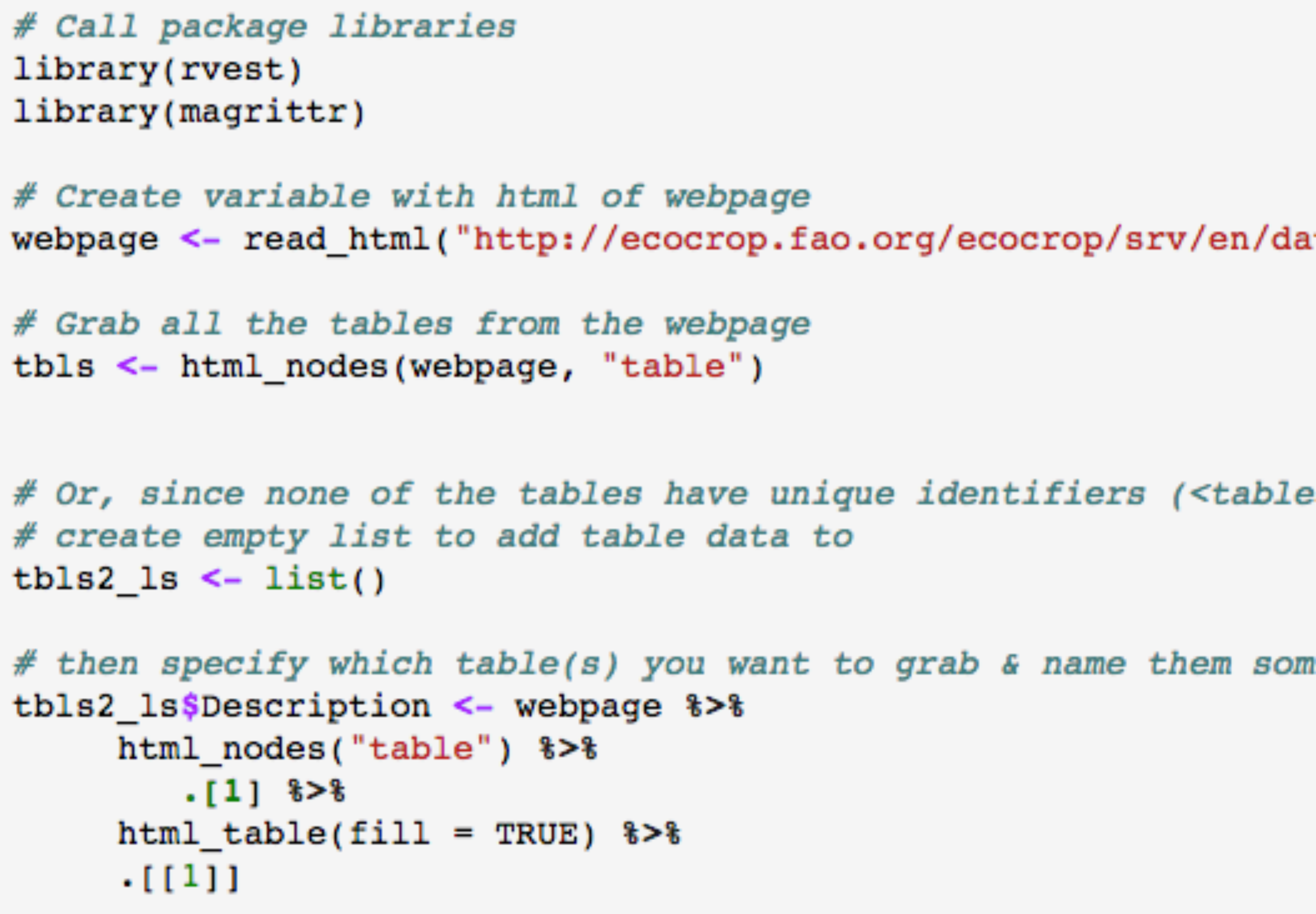


## Acacia decurrens

| Description | | | |
|---|---|---|---|
| **Life form** | tree | **Physiology** | evergreen, single stem |
| **Habit** | erect | **Category** | forest/wood, weed |
| **Life span** | perennial | **Plant attributes** | |

# Tables in HTML

- **Webscraping can help extract from html tables**
- rvest is a good package for this



```
# Call package libraries
library(rvest)
library(magrittr)

# Create variable with html of webpage
webpage <- read_html("http://ecocrop.fao.org/ecocrop/srv/en/da

# Grab all the tables from the webpage
tbls <- html_nodes(webpage, "table")


# Or, since none of the tables have unique identifiers (<table
# create empty list to add table data to
tbls2_ls <- list()

# then specify which table(s) you want to grab & name them som
tbls2_ls$Description <- webpage %>%
     html_nodes("table") %>%
        .[1] %>%
     html_table(fill = TRUE) %>%
     .[[1]]
```

The image part with relationship ID rId3 was not found in the file.

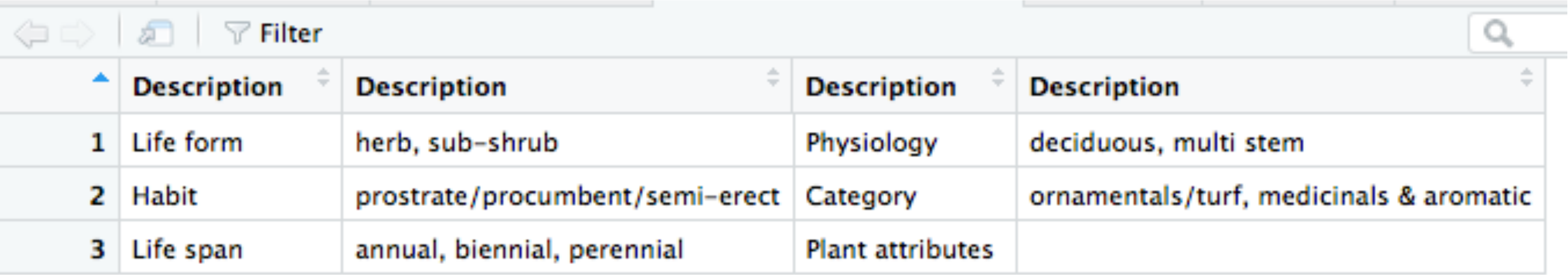
# Tables in R

- **SO MANY KINDS OF DATA:**
- Vectors. a <- c(1,2,5.3,6,-2,4) # numeric vector. ...
- Matrices. All columns in a matrix must have the same mode(numeric, character, etc.) ...
- Arrays. ...
- Data Frames. ...
- Lists. ...
- Factors. ...
- Data Frames

The image part with relationship ID rId3 was not found in the file.

# Tables in R

- Scraped html tables can be stored as various data types in R
- Getting them to compile into a single table require unknowable volumes of magic



| | Description | Description | Description | Description |
|---|---|---|---|---|
| 1 | Life form | herb, sub-shrub | Physiology | deciduous, multi stem |
| 2 | Habit | prostrate/procumbent/semi-erect | Category | ornamentals/turf, medicinals & aromatic |
| 3 | Life span | annual, biennial, perennial | Plant attributes | |

The image part with relationship ID rId3 was not found in the file.

# In Summary

- Damn near everything can be rendered into a table format, including tables
- Tables are everywhere
- You might be a table
- If not, R can help you wish you were

The image part with relationship ID rId3 was not found in the file.

# Pau

The image part with relationship ID rId3 was not found in the file.