

# The need for the R function Self\_Cleaning(data)

Marianna Simões  
University of Kansas

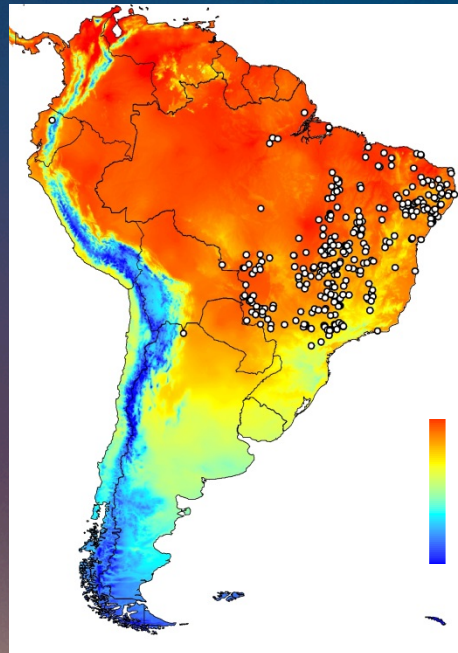
# Outline

---

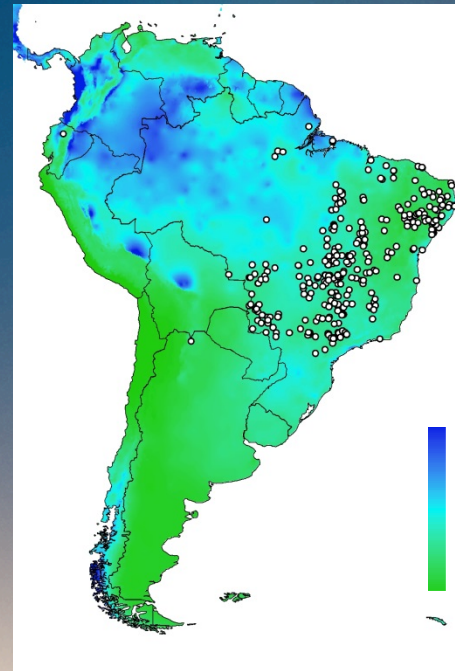
1. The Importance of Primary data
2. Associated Problems
3. Importance of : Data Cleaning & Specialists
4. Novel problems and new ways to clean the data
5. Ways to improve

# Primary Data

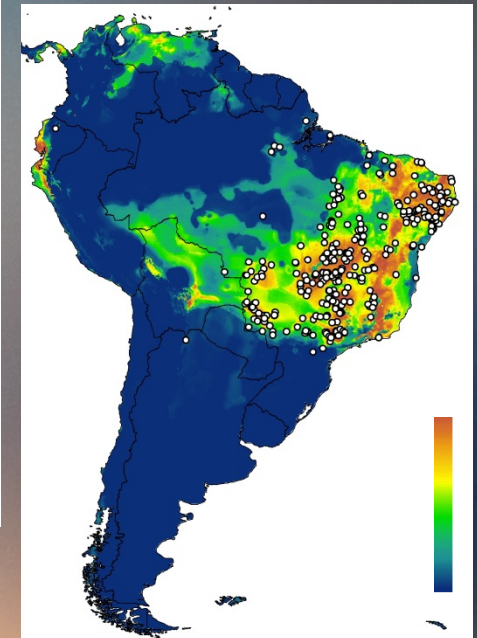
Observation of presences and absences of species through time and space.



Temperature



Precipitation



Niche Modeling



# Data Problems

---

- Misidentifications;
- Outdated taxonomy;
- Mislabeled specimen;
- Misspelling of taxa name, locality...
- Faulty georeferencing

Data Cleaning  
+  
Specialist

# Data Cleaning

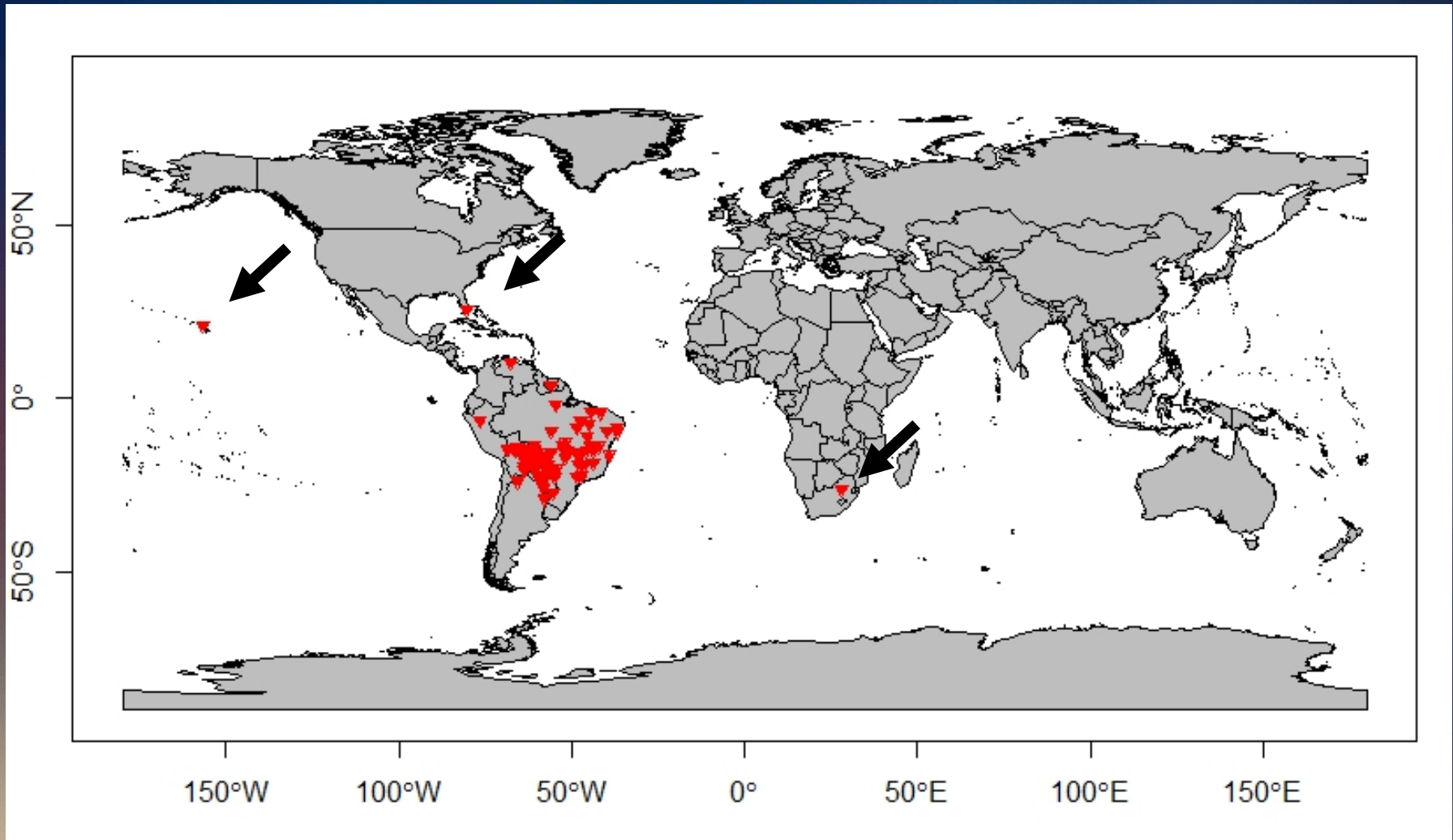
- Example 1 - *Tabebuia aurea*



# Data Cleaning

• Example 1- *Tabebuia aurea*

132 occurrence points



# Data Cleaning

- Example 2 - *Lecythis pisonis*







# The Specialist



(Ghostbusters, 1984)

# The Specialist

*Mesomphalia turrita* Boheman, 1850

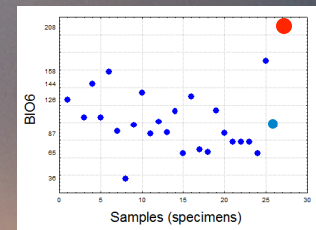
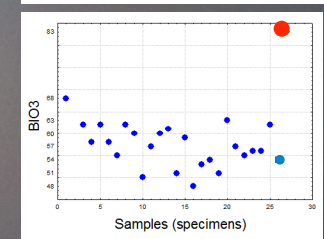
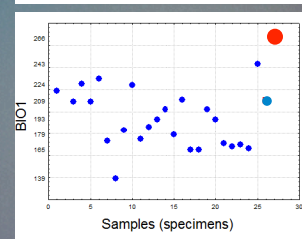
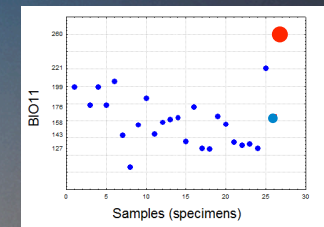
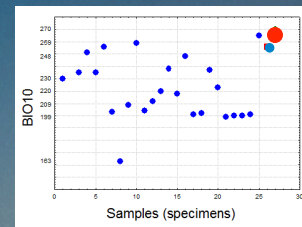
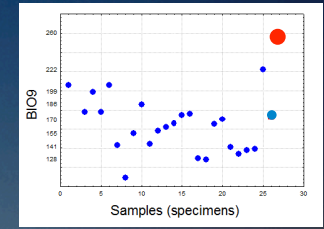
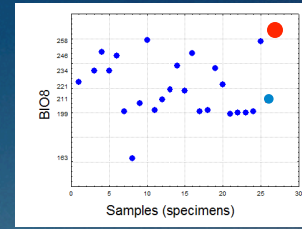
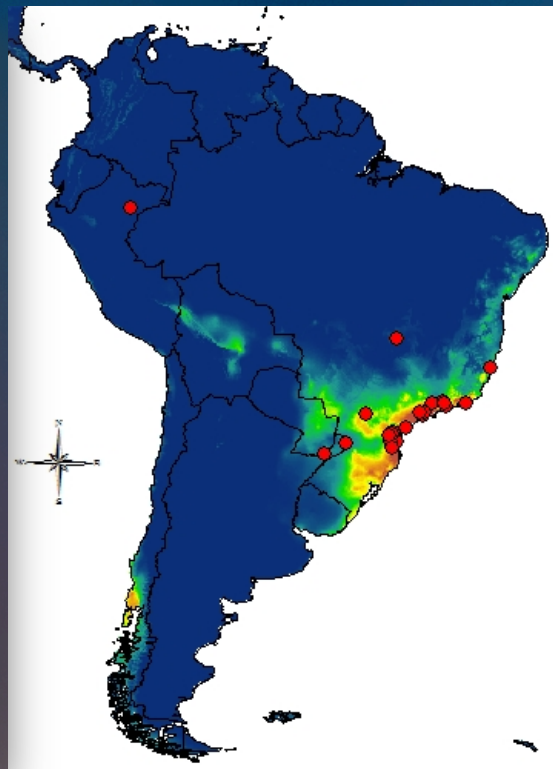
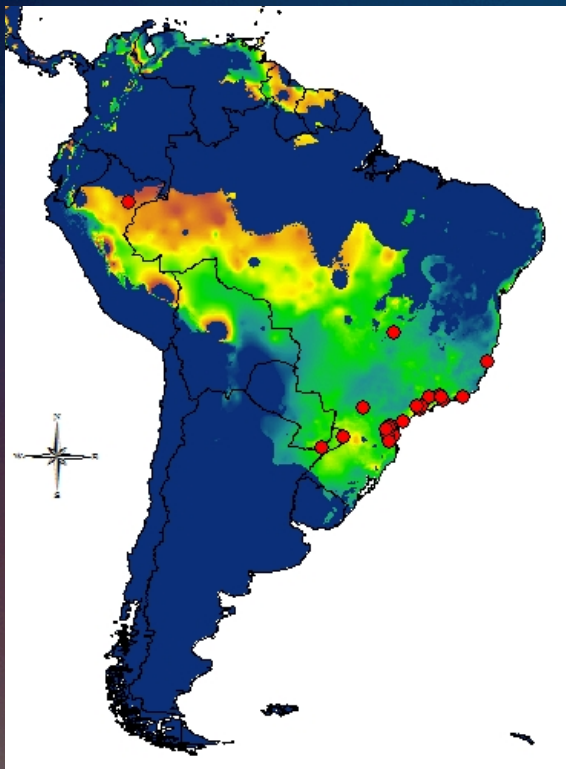
Borowiec & Takizawa, 2013



Label: [Loreto] Iquitos, Nanay, Nauta,  
I-IV.2003, Jose & Arai leg.

# The Specialist

*Mesomphalia turrata* Boheman, 1850



MISLABELED SPECIMEN!!!

# Ways to Improve

---

We have advanced a lot, ...but there's still a long way to go

- Database with photographs;
- Database with detailed information;
  - Number of records;
  - Number records georeferenced or not;
  - Collecting event date;
  - Collector

# Ways to Improve

species link  
desde 2002


ors dataCleaning tools tips preferences


### Search form ✕ close

any field

general	taxonomy	collect
barcode <input type="text"/>	determinator <input type="text"/>	collector <input type="text"/>
collection code <input type="text"/>	kingdom <input type="text"/>	collect number <input type="text"/>
catalog number <input type="text"/>	phylo <input type="text"/>	collect year <input type="text"/>
institution code <input type="text"/>	class <input type="text"/>	country <input type="text"/>
	order <input type="text"/>	state <input type="text"/>
	family <input type="text"/>	county <input type="text"/>
		locality <input type="text"/>

scientific name   clear

 phonetic search (scientific names)

search also for synonyms defined in the following dictionaries

- species2000 Catalog of Life
- List of Species of the Brazilian Flora
- Moure's Bee Catalogue
- DSMZ Bacteria Names

**search**

### search only within records

images

with images  without images

live material  pollen

material type

type  not type

red list

spp. in MMA red lists

spp. that are not red lists

geographic coordinates

with coords.  without coords.

original  by county

blocked by collection

coordinates quality

not suspect  are suspect

**gistros: Leia mais...**

**assista ao vídeo**  
**com dicas de uso da**  
**de speciesLink**

**YouTube**

**conheça as normas**

# Ways to Improve

species link  
since 2002

showing records from 1 to 100 of 109 for

**Attention!**  
The scientific names and genera listed at right are compared with available dictionaries according to the biological group. In **bold green** are the accepted ones, in **bold gray** the synonyms and in **orange** the ones that were not found. Family names are only checked for existing or not in the dictionaries. In the species inventory, the name appears in **blue** when the specimen is identified to the genus level only. See **tips** for detailed information.

Inventory of the records found

species <sup>1 2</sup>

<i>Dorynota</i>	2
<i>Dorynota (akantaka) aeneocincta</i>	1
<i>Dorynota (akantaka) bivittipennis</i>	6
<i>Dorynota (akantaka) kiesenwetteri</i>	3
<i>Dorynota (akantaka) peregrina</i>	3
<i>Dorynota (akantaka) tenebrosa</i>	4
<i>Dorynota (akantaka) truncata</i>	2
<i>Dorynota (akantaka) viridisignata</i>	9
<i>Dorynota</i> (s. str.) <i>bidens</i>	10
<i>Dorynota</i> (s. str.) <i>cornigera</i>	9
<i>Dorynota</i> (s. str.) <i>monoceros</i>	4
<i>Dorynota</i> (s. str.) <i>pugionata</i>	56

summary	names	records
<a href="#">Genus id only</a>	1	2
not found	11	107
<b>Total</b>	<b>12</b>	<b>109</b>

<sup>1</sup> Case insensitive  
<sup>2</sup> Names in **bold green** appear in the available dictionaries, in **bold gray** the synonyms and in **orange** the not found ones. When the identification is at genus level only, the name appears in **blue** in this inventory and is not checked against the available dictionaries. See "tips" for more information.

**ANIMALIA ARTHROPODA INSECTA COLEOPTERA CHRYSOMELIDAE**  
*Dorynota (akantaka) viridisignata* Boheman  
**DZUP-Coleoptera 87894.** BLOQUEADO Tocantins, Brasil, **01/06/1979.**  
© Coleção Entomológica Pe. Jesus Santiago Moure (Coleoptera) (DZUP-Coleoptera)

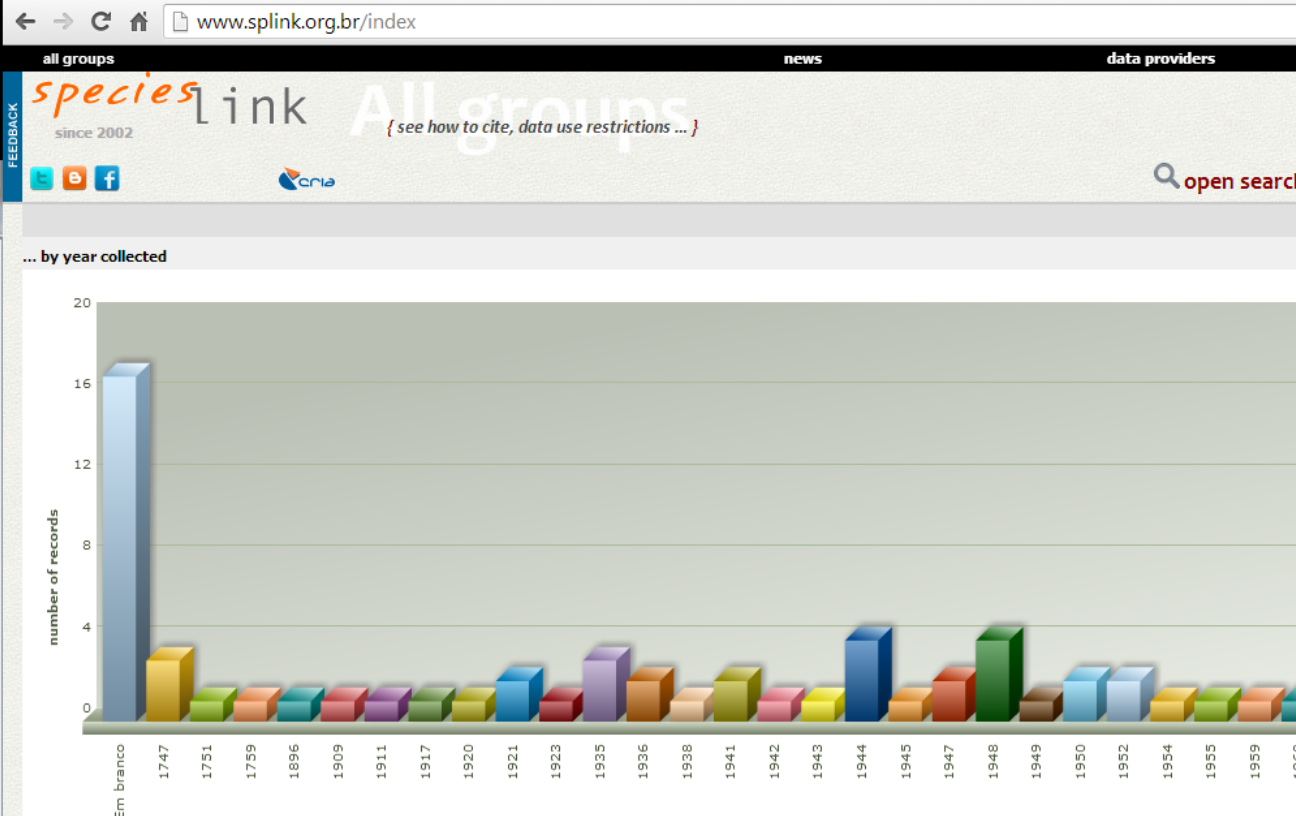
**ANIMALIA ARTHROPODA INSECTA COLEOPTERA CHRYSOMELIDAE**  
*Dorynota (akantaka) viridisignata* Boheman  
**DZUP-Coleoptera 87888** Col: R. Malkin BLOQUEADO, Mato Grosso, Brasil, **01/11/1964.**  
© Coleção Entomológica Pe. Jesus Santiago Moure (Coleoptera) (DZUP-Coleoptera)

**ANIMALIA ARTHROPODA INSECTA COLEOPTERA CHRYSOMELIDAE**  
*Dorynota* (s. str.) *pugionata* Germar  
**DZUP-Coleoptera 87769** BLOQUEADO, São Paulo, Brasil, **01/02/1759.**  
© Coleção Entomológica Pe. Jesus Santiago Moure (Coleoptera) (DZUP-Coleoptera)

**ANIMALIA ARTHROPODA INSECTA COLEOPTERA CHRYSOMELIDAE**  
*Dorynota* (s. str.) *pugionata* Germar  
**DZUP-Coleoptera 87768** Col: L. Stowaunenko. BLOQUEADO, BLOQUEADO, São Paulo, Brasil, **01/02/1962.**  
© Coleção Entomológica Pe. Jesus Santiago Moure (Coleoptera) (DZUP-Coleoptera)

**ANIMALIA ARTHROPODA INSECTA COLEOPTERA CHRYSOMELIDAE**  
*Dorynota (akantaka) aeneocincta* Spaeth  
**DZUP-Coleoptera 87741** BLOQUEADO, Amazonas, Brasil, **01/04/1935.**  
© Coleção Entomológica Pe. Jesus Santiago Moure (Coleoptera) (DZUP-Coleoptera)

# Ways to Improve



years	
<i>Em branco</i>	17
1747	3
1751	1
1759	1
1896	1
1909	1
1911	1
1917	1
1920	1
1921	2
1923	1
1935	3
1936	2
1938	1
1941	2
1942	1
1943	1
1944	4
1945	1
1947	2
1948	4
1949	1
1950	2
1952	2



# Ways to Improve

## Data summary

The graphs and information below are a brief analysis of the data retrieved in the last search.

coordinates	records
by municipality	297
original	397
blocked	1
without coordinates	230

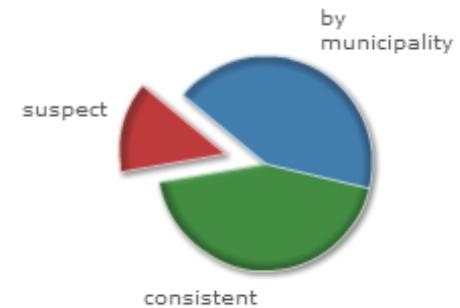
  

suspect	95
consistent	302
by municipality	297

### coordinates origin



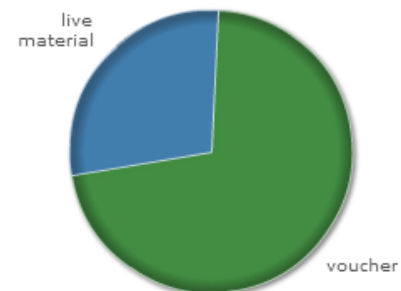
### coordinates quality



considers this information sensitive. **Suspect coordinates** refer to records with original coordinates that do not fall within the municipalit

images	quantity
voucher	20
live material	8
pollen	0
total	28

### type of material with images



# Ways to Improve

speciesLink Network | dataCleaning | lecythis pisonis map distrib... | The Short Lab

www.splink.org.br | cnpq







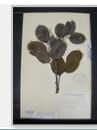























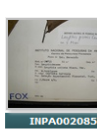



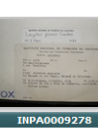
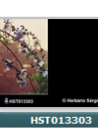
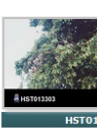
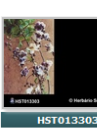

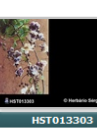








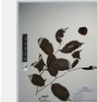



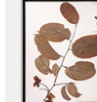


todos os grupos | noticias | provedores | indicadores | dataCleaning | ferramentas | dicas de uso | preferências

species link Todos os grupos  
desde 2002 { ver como citar, condições de uso dos dados ... }

abrir formulário de busca

resumo | imagens | mapa | gráfico | download

mostrando imagens de 1 a 100 das 172 encontradas

 NY00689572	 NY00689529	 INPA0044069	 INPA0044069	 INPA0044069	 NY00689598	 INPA0047235	 INPA0047235	 INPA0047235	 NY00689546	 NY00689539
 NY00389909	 NY00689584	 INPA0032946	 INPA0032946	 INPA0032946	 INPA0234347	 INPA0234347	 INPA0161763	 INPA0161763	 INPA0161763	 NY00689518
 UEC095809	 NY00689591	 NY00389907	 NY00689525	 NY00689553	 NY00689556	 NY00389916	 INPA0020851	 INPA0020851	 HUEF0003691	 HUEF00109465
 INPA0009278	 INPA0009278	 HST013303	 HST013303	 HST013303	 HST013303	 HST013303	 INPA0116598	 INPA0116598	 NY00689545	 NY00689521
										

# Conclusion

---

Importance of Data Cleaning

Role of the Specialist

# Thank you!

