

# Refinamento de Dados

Qualidade das informações biológicas digitalizadas no SiBBR

---

Laura Rocha Prado

Museu de Zoologia da Universidade de São Paulo

1. Errar é humano
2. Como evitar erros
3. Como corrigir erros e padronizar dados

Errar é humano

---

*Errare humanum est, sed in errare perseverare diabolicum.*

Sêneca

## Protocolo de Digitalização

### 1. Ler etiquetas

7.11.55  
Barnier

MZ010745

## Protocolo de Digitalização

1. Ler etiquetas
2. Digitar informações no Excel

## Protocolo de Digitalização

1. Ler etiquetas
2. Digitar informações no Excel
3. Subir informações para o Specify



- Como o computador sabe o que é um erro?

- Como o computador sabe o que é um erro?
- Ele não sabe!

- Como o computador sabe o que é um erro?
- Ele não sabe!
- O computador aceita tudo.

- Como o computador sabe o que é um erro?
- Ele não sabe!
- O computador aceita tudo.
- Dados errados são arquivados como se fossem corretos.

- Como o computador sabe o que é um erro?
- Ele não sabe!
- O computador aceita tudo.
- Dados errados são arquivados como se fossem corretos.
- Efeito nas buscas.

- Como o computador sabe o que é um erro?
- Ele não sabe!
- O computador aceita tudo.
- Dados errados são arquivados como se fossem corretos.
- Efeito nas buscas.
- Efeito nas árvores taxonômicas e geográficas.

- Como o computador sabe o que é um erro?
- Ele não sabe!
- O computador aceita tudo.
- Dados errados são arquivados como se fossem corretos.
- Efeito nas buscas.
- Efeito nas árvores taxonômicas e geográficas.
- Atualmente a única maneira de corrigir os dados que já foram pro Specify é manualmente, dado por dado.

- Como o computador sabe o que é um erro?
- Ele não sabe!
- O computador aceita tudo.
- Dados errados são arquivados como se fossem corretos.
- Efeito nas buscas.
- Efeito nas árvores taxonômicas e geográficas.
- Atualmente a única maneira de corrigir os dados que já foram pro Specify é manualmente, dado por dado.
- 2000 dados errados = 2000 dados corrigidos manualmente



## Como evitar erros

---

- Pesquisa simples na internet

- Pesquisa simples na internet
- Pesquisa em catálogos taxonômicos

# Confirmação dos dados

- Pesquisa simples na internet
- Pesquisa em catálogos taxonômicos
- Pesquisa em bancos de dados geográficos/mapas

# Confirmação dos dados

- Pesquisa simples na internet
- Pesquisa em catálogos taxonômicos
- Pesquisa em bancos de dados geográficos/mapas
- Tem dúvida em relação a qual dado entra em qual coluna?  
Pergunte!

# Confirmação dos dados

- Pesquisa simples na internet
- Pesquisa em catálogos taxonômicos
- Pesquisa em bancos de dados geográficos/mapas
- Tem dúvida em relação a qual dado entra em qual coluna? Pergunte!
- Dados que não se encaixam em nada? Use colunas de comentários (Remarks)!

## Práticas para numeração e cópias

- Selecionar várias linhas e arrastar para criar números em série
- Selecionar uma linha e arrastar para copiar valores idênticos

## Fórmulas

- Concatenação: união de campos separados
- Contagem: soma de valores de colunas (número de exemplares)
- Cópia: para valores repetidos (altitude, data)
- Testes (IF Statements): para copiar valores variáveis (localidade)

## Como corrigir erros e padronizar dados

---



Antigamente conhecido como **Google Refine**, é uma ferramenta que, como o próprio nome diz, é usada para refinar dados em tabelas extensas. O **Open Refine** deve ser usado **antes** da importação dos dados para o **Specify**.



1. Entrar em <http://openrefine.org/download.html>
2. Baixar o pacote apropriado (normalmente: Windows kit, versão 2.7)
3. Descomprimir o arquivo .zip na pasta desejada

## Pré-requisito

- A máquina precisa ter o Java Runtime Environment instalado.
- Verificar JRE: `cmd -> java -version`
- Se não tiver JRE, baixar em <https://www.java.com/en/download/>

# Iniciando/Finalizando o Open Refine

1. *Iniciar* o servidor antes de acessar a página do Open Refine
2. *Finalizar* o servidor quando desejar fechar a aplicação

## Iniciar

- Vá até a pasta onde o Open Refine foi extraído
- Clique duas vezes em OpenRefine.exe
- O site do OpenRefine deve ser aberto automaticamente (ou vá para <http://127.0.0.1:3333/>)

## Finalizar

- Vá até a janela de comando aberta pelo Open Refine
- Aperte as teclas CTRL+C
- Espere até que a janela feche sozinha
- Se aparecer a pergunta "Terminate all batch processes? Y/N", aperte Y e aguarde

**Facets** categorias de dados

**Cluster** agrupamento de dados usando diferentes algoritmos

Dados para padronizar ou corrigir: coletor, táxon, localidade

# Bônus! Reconciliação de dados

- Wikidata
- Use o serviço de reconciliação do Wikidata para verificar nomes de táxons e localidades!

- Erros podem acontecer em todas as fases do processo de digitalização;

- Erros podem acontecer em todas as fases do processo de digitalização;
- Felizmente há como evitar e como corrigir esses erros;

- Erros podem acontecer em todas as fases do processo de digitalização;
- Felizmente há como evitar e como corrigir esses erros;
- Use métodos no Excel para evitar erros;



- Erros podem acontecer em todas as fases do processo de digitalização;
- Felizmente há como evitar e como corrigir esses erros;
- Use métodos no Excel para evitar erros;
- Use o Open Refine para corrigir erros e padronizar dados;

- Erros podem acontecer em todas as fases do processo de digitalização;
- Felizmente há como evitar e como corrigir esses erros;
- Use métodos no Excel para evitar erros;
- Use o Open Refine para corrigir erros e padronizar dados;
- **Limpe os dados antes de fazer a importação no Specify;**

- Erros podem acontecer em todas as fases do processo de digitalização;
- Felizmente há como evitar e como corrigir esses erros;
- Use métodos no Excel para evitar erros;
- Use o Open Refine para corrigir erros e padronizar dados;
- Limpe os dados antes de fazer a importação no Specify;
- Evite o trabalho de correção posterior;

- Erros podem acontecer em todas as fases do processo de digitalização;
- Felizmente há como evitar e como corrigir esses erros;
- Use métodos no Excel para evitar erros;
- Use o Open Refine para corrigir erros e padronizar dados;
- Limpe os dados antes de fazer a importação no Specify;
- Evite o trabalho de correção posterior;
- **Padronize os dados para manter a qualidade de informação.**



D. Huynh.

**Open Refine.**

<http://openrefine.org/>, 2017.



W. Labs.

**OpenRefine-Wikidata interface.**

<https://tools.wmflabs.org/openrefine-wikidata/>.



R. Page.

**Using Google Refine and taxonomic databases (EOL, NCBI, uBio, WORMS) to clean messy data.**

<http://iphylo.blogspot.com.br/2012/02/using-google-refine-and-taxonomic.html>, 2012.



R. Page.

**Surfacing the deep data of taxonomy.**

*ZooKeys*, 550:247–260, 2016.



L. R. Prado.

**SiBBR - MZUSP - Entomologia.**

[https://arbolitoloco.github.io/sibbr\\_mzusp/](https://arbolitoloco.github.io/sibbr_mzusp/), 2017.

# Roube essa apresentação

Baixe esses slides e um arquivo de texto que preparei com cada passo-a-passo em

<http://bit.do/sibbr>

feito com  $\text{\LaTeX}$

Creative Commons Attribution-ShareAlike 4.0 International License.



Dúvidas?