1 Errar é humano

Errare humanum est, sed in errare perseverare diabolicum.
Sêneca

A frase latina, atribuída ao filósofo grego Sêneca, diz: Errar é humano, porém persistir no erro, por arrogância, é diabólico. Humanos cometem erros o tempo todo. Infelizmente, os tão naturais erros humanos prejudicam imensamente a análise de dados coletados em grande volume. Erros, quando existentes em um conjunto de dados muito grande, são quase impossíveis de serem detectados e corrigidos. No projeto de digitalização de espécimes biológicos do qual o Museu de Zoologia da Universidade de São Paulo faz parte, milhares de registros são gerados mensalmente, e poucas pessoas estão disponíveis para avaliar a qualidade desses dados. Apesar de não ser possível eliminar os erros, felizmente é possível evitá-los, detectá-los em uma escala menor, e corrigí-los antes que eles sejam importados para o banco de dados permanentemente. Para manter um nível adequado de qualidade e confiabilidade dos dados digitalizados, é imprescindível que as planilhas geradas passem por algum protocolo de refinamento de dados.

1.1 Protocolo de digitalização

O processo de digitalização dos exemplares do MZUSP passa por várias fases[5]:

1. Separação das etiquetas dos exemplares
2. Fotografia das etiquetas
3. Leitura das etiquetas
4. Transcrição do texto das etiquetas para a planilha
5. Importação da planilha para o Specify

Em todas as fases do protocolo é possível cometer erros de diversos tipos. A seguir, identificam-se os erros mais comuns.

1.2 Erro de etiquetas

Esse é o mais grave erro que pode ser cometido durante a digitalização dos espécimes. Ao separar as etiquetas dos exemplares para a posterior fotografia, deve-se prestar muita atenção à quais etiquetas pertencem a qual exemplar. Uma associação errônea da etiqueta a seu exemplar de origem invalida toda a informação científica relevante.

1.3 Erros de leitura

Erros de leitura normalmente podem ser atribuídos a má-impressão ou caligrafia dúbia das informações na etiqueta, ou à falta de experiência do agente digitalizador, que não reconhece os nomes ali escritos.

1.4 Erros de digitação

Durante a transcrição dos dados de cada etiqueta, podem acontecer muitos erros de digitação. Esse é o tipo mais comum de erro e, felizmente, um dos tipos mais facilmente solucionáveis.

1.5 Erros de categoria de dados

A planilha de dados utilizada no projeto é padronizada para cada coleção, de modo que cada coluna deve conter um tipo específico de informação. Durante o processo de importação dos dados, as colunas do banco de dados são associadas às colunas equivalentes na planilha. Um erro comum desse tipo é o de preencher informações em colunas erradas. Como o Specify não consegue validar os dados em todas as colunas da planilha importada, erros desse tipo geralmente são armazenados no banco de dados sem serem detectados. Esse é um erro muito grave e de difícil correção, uma vez que descaracterizam as informações armazenadas.

2 O problema dos erros: efeito-cascata

O principal resultado dos erros humanos em bancos de dados biológicos é a deterioração das informações armazenadas e sua potencial invalidez. A diminuição do erro humano resulta diretamente no aumento da confiabilidade dos dados. Se o banco de dados armazena informações equivocadas, diversas consequências podem ser observadas.

2.1 Efeitos nas consultas de dados

A utilidade de um banco de dados está relacionada à capacidade de sua manipulação. Precisamos extrair informações específicas de um conjunto de dados muito grande. Para isso, utilizamos as consultas, ou *queries*. Quando um erro é armazenado no banco de dados, as consultas retornarão resultados incompletos ou equivocados. Pior, é possível que não encontremos quaisquer resultados. A informação armazenada ali, então, torna-se inútil. Imagine que você se cadastrou para concorrer a um sorteio de um milhão de reais em barras de ouro (que valem mais do que dinheiro). Você usou seu CPF como identificador do seu cadastro, mas, sem querer, informou seu telefone erroneamente, invertendo os dois últimos dígitos. O que acontecerá se você for sorteado como vencedor? Ficará de mãos vazias, porque não será possível contatá-lo para informar do grande prêmio. Acontece algo similar com os dados biológicos que são armazenados no SiBBr. Exemplares serão virtualmente perdidos, porque não será possível encontrá-los a partir das informações existentes.

2.2 Efeitos nas árvores taxonômicas e geográficas

O Specify utiliza dados taxonômicos e geográficos para atualizar as suas árvores hierárquicas. Como não há validação desses dados, qualquer erro de digitação é tratado como um registro novo. O resultado é uma série de hierarquias caóticas, que não refletem a realidade dos dados. Espécimes passam, então, a "desaparecer"das classificações biológicas e dos mapas elaborados a partir dos dados geográficos associados.

3 Como evitar erros

Felizmente os erros humanos podem ser prevenidos, corrigidos e padronizados. Consideramos, a seguir, alguns métodos de refinamento de dados que podem ser úteis.

3.1 Pesquisa prévia para confirmação de dados

Recomenda-se a pesquisa das informações dúbias para confirmação, especialmente dos nomes dos táxons e das localidades geográficas. Para os táxons, recomenda-se que os nomes sejam confirmados em catálogos e/ou listas de espécies, e sua classificação taxonômica seja atualizada (por exemplo Família, Tribo, ou outro nível hierárquico adequado). Para as localidades, recomenda-se que a sua existência seja confirmada a partir da pesquisa em bancos de dados específicos, mapas ou *gazeteers*. A seção de bônus indica uma maneira de auxiliar a verificação desses dados.

3.2 Adequação de dados na planilha

Cada coluna da planilha de dados categoriza um tipo particular de informação. Por isso, é necessário ter conhecimento prévio de como a planilha foi montada para cada organismo com o qual se está trabalhando. Recomenda-se, sempre que houver dúvida, acompanhar a explicação das colunas das tabelas, disponível no repositório para o SiBBr/Entomologia (`http://bit.do/sibbr`), ou perguntar ao seu supervisor/curador o que deve ser feito. Por favor não adicione informações em colunas sem antes confirmar se o lugar delas está correto. Como já discutido anteriormente, esse é um dos erros mais graves que pode ser cometido.

3.3 Uso de atalhos no Excel

Certos dados devem ser digitados mais de uma vez na planilha. Para facilitar e garantir que os dados estarão corretos, algumas estratégias podem ser utilizadas.

3.3.1 Números em sequência

Por exemplo, para números em sequência, recomenda-se que as células anteriores sejam selecionadas e arrastadas para as próximas células vazias, no lugar de digitar os números. Assim, erros de digitação são evitados.

3.3.2 Nomes/textos iguais

Alguns campos utilizam dados que serão repetidos. Por exemplo, quando há vários exemplares de uma mesma espécie, ou quando só há uma data de coleta (ou seja, a data de coleta inicial será idêntica à final). Nesse caso, é melhor selecionar o dado e arrastá-los para repetir, ou usar o atalho CTRL+C, ou então usar a fórmula:

```
=CÉLULACOPIADA
```

onde o texto deve ser substituído pelo código da célula a ser copiada. Por exemplo, se a data inicial estiver na célula B1, colocando-se a fórmula `=B1` copia-se a data inicial para a data final. Assim evitam-se erros de digitação.

3.3.3 Preenchimento condicional

Certos campos não podem estar vazios na planilha. Durante a importação da planilha para o Specify, a presença de certos dados vazios resultará em erros. É o caso da coluna *LocalityName*, por exemplo, que deve repetir o último dado geográfico disponível (se houver). Para evitar a correção desses dados apenas durante a importação, é possível usar a opção de

preenchimento condicional no Excel. Nessa opção, podemos indicar que, se a coluna anterior estiver preenchida, então a próxima deve repetir aquele dado. Por exemplo: se B1 estiver vazia, B2 também será vazia. Em caso contrário, B2 deve estar preenchida com o mesmo valor. Um exemplo de fórmula seria: Um exemplo de fórmula seria: Um exemplo de fórmula seria: Um exemplo de fórmula seria: Um exemplo de fórmula seria: Um exemplo de fórmula seria: Um exemplo de fórmula seria: Um exemplo de fórmula seria:

```
=SE(ÉCÉL.VAZIA(B1);"";B1)
```

4 Como corrigir e padronizar erros

Os tópicos a seguir discutem na prática como usar o Open Refine para refinar os dados contidos nas planilhas a serem importadas pelo Specify.

4.1 O que é o Open Refine

Open Refine é uma ferramenta gratuita e *open source*, poderosa para trabalhar com dados bagunçados, melhorando-os[1]. Inicialmente difundida como Google Refine, essa ferramenta analisa conjunto de dados e permite o seu refinamento facilmente.

4.2 Instalação

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. Para verificar se o JRE está instalado, entre no *prompt* de comando do Windows (tecla Windows -¿ digitar cmd -¿ tecla Enter) e digite:

```
java -version
```

Aperte a tecla Enter. No caso de aparecer a mensagem ”Java command not found”, baixe o JRE acessando https://www.java.com/en/download/. Instale o pacote conforme indicado no *site*.

4.3 Inicialização/Finalização

O Open Refine funciona de maneira um pouco diferente dos outros programas convencionais. Ele funciona como uma página da web, que está hospedada dentro de um servidor local dentro do seu próprio computador. Por isso você precisa:

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

4.3.1 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/)

4.3.2 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

4.4 Criar novo projeto

1. Para começar a refinar uma planilha, clique, no menu lateral esquerdo, em ”Create Project”. Para continuar com projetos abertos anteriormente, clique em ”Open Project”.
2. No menu aberto, aparecerá a opção para importar dados (”Get data from this computer”). Selecione a planilha que deseja importar (arquivos .xls, .xlsx ou .csv).
3. Clique em ”Next”.
4. Aguarde que a prévia apareça.
5. Na caixa superior direita digite o nome que deseja dar ao projeto.
6. Clique em ”Create Project”.

4.5 Métodos de refinamento

O Open Refine mostra até 50 linhas da planilha por tela. Para agrupar os dados e manipulá-los, escolha uma coluna. Por exemplo, a coluna de Coletores (Collectors).

4.5.1 Facets

Para agrupar os dados, o Open Refine usa algo chamado ”Facets”(facetas). Essas facetas são nada mais do que categorias de alto nível.

1. Vá até a coluna de interesse. Por exemplo: Collectors.
2. Clique na seta para baixo que aparece ao lado do nome da coluna.
3. Clique em ”Facet”.
4. Clique em ”Text facet”.

Uma janela aparecerá à esquerda, listando os resultados repetidos. Avaliando a lista de nomes, é possível rapidamente encontrar nomes com erros de digitação, por exemplo. O mesmo nome escrito de mais de uma maneira diferente. Ao lado de cada nome, aparece um número, que indica a quantidade de registros associadas a ele. Ao clicar em um nome qualquer, a tabela é filtrada e mostra apenas os registros associados. Para cada coluna existente é possível criar facetas e exibir os resultados. Assim, facetas devem ser criadas para cada coluna que se deseje editar.

4.5.2 Cluster

Ao criar facetas, é possível forçar o agrupamento de dados, usando a opção de ”Cluster”do Open Refine. É necessário, então, selecionar o(s) algoritmo(s) que realizarão a comparação entre os dados e escolher quais devem ou não ser agrupados.

1. Na janela de facetas, clique no botão ”Cluster”.
2. Selecione as opções de método (”Method”) e função (”Keying function”) de modo a verificar quais agrupamentos são formados.
3. Se um grupo for formado, as opções agrupadas aparecerão listadas uma acima da outra. Clique em ”Merge”para mesclar as opções que parecem iguais, e selecione qual texto deve aparecer como o selecionado.
4. Clique em ”Merge Selected & Re-Cluster”.

Esse procedimento pode ser realizado quantas vezes for necessário, para cada coluna desejada.

4.5.3 Exportar

Para salvar o projeto e fazer a importação para o Specify, clique no botão ”Export”, no canto superior direito e salve o arquivo para o formato Excel (.xls).

5 Bônus: outras utilidades

O Open Refine é uma ferramenta muito versátil e poderosa, e sua utilidade vai além da padronização de dados local. Há uma categoria de refinamento de dados, que utiliza serviços de reconciliamento de dados externos para comparação e limpeza de dados da planilha. Esses serviços são fornecidos por pessoas ou organizações variados[3, 4]. Vamos mostrar dois exemplos de comparação de dados que podem ser utilizados com o serviço de reconciliação de dados do Wikidata[2].

5.0.1 Verificação de nomes taxonômicos e geográficos

Para verificar se os nomes dos táxons foram inseridos corretamente na coluna, crie na tabela uma coluna que concatena (une) o nome do gênero e o nome da espécie.

1. Crie um novo projeto no OR seguindo o protocolo tradicional.
2. Clique na coluna do nome completo do táxon e selecione a opção ”Reconcile”e ”Start reconciling”.
3. Na janela que aparecerá, clique no botão inferior à esquerda ”Add Standard Service”.
4. Adicione o serviço de reconciliação desejado. No nosso caso, o endereço do serviço do Wikidata (https://tools.wmflabs.org/openrefine-wikidata/en/api).
5. Na próxima janela, na coluna da esquerda, selecione o dado do Wikidata que combina com o dado que você quer comparar. No nosso caso, selecione a primeira opção, ”taxon”.
6. Clique no botão ”Start reconciling”, no canto inferior direito.
7. Aguarde o processamento dos dados.

Quando o nome é marcado em azul o serviço encontrou dados que combinam com aquela informação, isto é, o nome foi verificado. Se aparecem vários nomes abaixo da informação original, significa que vários nomes foram parcialmente encontrados combinando. Se um desses nomes for o mais adequado, clique na seta única para indicar que o nome foi aceito. Quando nenhuma combinação é encontrada, é possível fazer uma busca, clicando-se em ”Search for match”. Ao se clicar no nome que contiver um link, a página é redirecionada para o registro do Wikidata que bate com aquela informação. Esses mesmos passos podem ser seguidos para informações geográficas. Basta criar um campo que junta todas as informações de localidades (unindo País, Estado, Cidade e Localidade, por exemplo), ou fazer a busca a partir de municípios ou estados.

6 Conclusão

O processo de digitalização de espécimes possui várias fases. Erros humanos podem ocorrer em todas as fases do processo. Felizmente há maneiras de corrigir a maior parte desses erros. Alguns procedimentos no Excel, como uso de fórmulas, evitam erros de digitação por repetição. Já o uso do Open Refine auxilia na detecção de erros de digitação, além de permitir a padronização e a checagem de dados de diversos tipos.

Referências

[1] David Huynh. *Open Refine*. 2017. URL: http://openrefine.org/.

[2] Wikimedia Labs. *OpenRefine-Wikidata interface*. URL: https://tools.wmflabs.org/openrefine-wikidata/ (acesso em 15/08/2017).

[3] Roderic Page. *Using Google Refine and taxonomic databases (EOL, NCBI, uBio, WORMS) to clean messy data*. personal blog. 2012. URL: http://iphylo.blogspot.com.br/2012/02/using-google-refine-and-taxonomic.html (acesso em 15/08/2017).

[4] Roderic Page. "Surfacing the deep data of taxonomy". Em: *ZooKeys* 550 (jan. de 2016), pp. 247–260. ISSN: 1313-2970, 1313-2989. DOI: 10.3897/zookeys.550.9293. URL: http://zookeys.pensoft.net/articles.php?id=6232.

[5] Laura Rocha Prado. *SiBBr - MZUSP - Entomologia*. 2017. URL: https://arbolitoloco.github.io/sibbr_mzusp/ (acesso em 15/08/2017).