Curso de formação
Publicação de dados de biodiversidade através do GBIF

Abril de 2020

Rui Figueira
Nó Português do GBIF, Instituto Superior de Agronomia
ruifigueira@isa.ulisboa.pt

Exercício Prático: OpenRefine

Objetivo

Realizar um conjunto de operações elementares de seleção, deteção e limpeza de dados utilizando OpenRefine.

Tarefa 1 - Instalação do OpenRefine (se necessário)

1. Aceda a http://openrefine.org
2. Descarregue o software adequado para o seu sistema operativo
3. Instale segundo as instruções disponíveis no site.

Tarefa 2. Criar o projecto

1. Descarregar o ficheiro de dados de exemplo em https://goo.gl/d1kA61
2. Abrir o OpenRefine e importar os dados
3. Dar um nome ao projeto e criá-lo.

Tarefa 3. Criar facetas / filtros e corrigir texto

1. Crie uma faceta de texto para cada uma das colunas: Collectors, País, Província
2. Identifique a existência de incoerências de valores com a função "**cluster**" e corrija-as
3. Para outros valores incoerentes, corrigir de forma massiva editando-os diretamente no filtro.

Tarefa 4. Identificação de valores duplicados

1. Para o campo Catalog Number, fazer a verificação de existência de valores duplicados, através da função **Facet → Customized facets → Duplicates facet**
2. Corrija os valores duplicados detetados (para efeitos de exemplo, atribua um número diferente, adicionando o sufixo "a"), e volte a verificar a existência de valores duplicados.

Tarefa 5. Criar a coluna eventDate - opção 1

Sabendo que o formato de data recomendado pelo DwC é AAAA-MM-DD, mas que o a coluna contém os dados no formato DD/MM/AAAA:

1. Separar os valores em três novas colunas, através da função **Edit column → Split into several columns**, usando o caracter "/" como separador de colunas
2. Criar uma nova coluna baseada na coluna Start Date 3, com a opção **Edit column → Add column based on this column...** , com o nome eventDate. Ao campo Expression, adicione

   ```
   cells["Start Date 3"].value + "-" + cells["Start Date 2"].value + "-" + cells["Start Date 1"].value
   ```

   Esta expressão junta os valores das três colunas Start Date 1, 2 e 3 utilizando o operador +. Clique OK.

Tarefa 6. Criar a coluna eventDate - opção 2

1. Antes de mais, desfazer as ações até à situação anterior à Tarefa 5. Usar a função **Undo / Redo**
2. A seguinte operação apenas pode ser aplicada sobre às datas completas (com o formato DD/MM/AAAA), pelo que é necessário criar um filtro isolar estes registos. Estes podem ser selecionados utilizando a coluna Start Date Precision, com o valor 1. Para isso aplicar uma faceta de texto à coluna Start Date Precision, e selecionar apenas os registos com valor 1.
3. Após selecionar os registos que têm apenas datas completas (Start Date Precision com o valor 1), transformar os valores em datas com a função **Edit cells → Common transforms → To date**
4. Selecionar, de seguida, apenas os registos com valores Start Date Precision igual a 2 (estes contêm na data valores com o formato MM/AAAA, que têm de ser transformados no formato AAAA-MM)
5. Fazer a transformação da coluna Start Date selecionando **Edit cells → Transform...** , e utilizando a expressão

   ```
   value.split("/")[1] + "-" + value.split("/")[0]
   ```

   Esta expressão junta através do operador "+" o resultado da aplicação da função *split* sobre o valor da célula, utilizando o separador "/". O valor entre parêntesis recto indica qual das partes resultantes do *split* é para reter. Clique OK
6. Remover o filtro.

Tarefa 7. Verificar e corrigir valores de latitude anómalos

1. Converter as colunas Latitude1 e Longitude1 para valores numéricos, selecionando **Edit cells → Common transforms → to Number**
2. Sabendo que a latitude de Moçambique varia aproximadamente -10º e -27º, verificar a existência de valores fora destes limites. Para isso, aplicar uma faceta numérica à coluna Latitude1, e usar os cursores da faceta para definir os limites que identifiquem valores anómalos: > -10. (Sugestão: aplicar um filtro sobre a coluna País para isolar apenas os registos de Moçambique)
3. Verificar que são selecionadas 16 linhas com valores de latitude anómalos para Moçambique. Sabendo que estes valores deveriam ser negativos (multiplicando por -1), fazer a correção selecionando **Edit cells → Transform**, e aplicar a expressão

   ```
   value * (-1)
   ```
4. Verificar que na faceta Latitude1, os valores aparecem agora dentro dos limites desejados.

Tarefa 8. Exportar o projeto e a tabela

1. Exportar o projeto com o botão Export no canto superior direito. Será produzido um ficheiro comprimido com o projeto completo, incluindo a tabela de dados e o script com as operações realizadas. Este projeto pode ser reaberto no futuro, e inclui todas as operações realizadas.
2. Exportar a tabela de dados no mesmo botão, para um dos formatos desejados: tabela tsv, tabela csv, HTML, Excel, tabela ODF.

Referências

http://openrefine.org

https://github.com/OpenRefine/OpenRefine/wiki

http://enipedia.tudelft.nl/wiki/OpenRefine_Tutorial