Curso de formação: Publicação de dados de Biodiversidade através do GBIF - 2022

Lisboa, 1-3 de Junho 2022

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

Exercício Prático 1: Boas práticas no uso de folhas de cálculo em manuseamento de dados

Objectivo

Identificar os erros que podem ocorrer na manipulação de dados através de uma folha de cálculo. Neste exercício serão utilizados dois programas, MS Excel e LibreOffice Calc, destacando-se as vantagens de cada um deles na manipulação de dados.

Preparação dos exercícios

Descarregue os ficheiro no seguinte link para uma directoria da sua área de trabalho.

Parte 1. Manipular a abertura e escrita de ficheiros

Tarefa 1 - Diferença entre Abrir e Importar para um ficheiro de texto (csv), em Excel

1. Abrir o ficheiro `dados_sampleQLDB_ex01_code01.csv` com o Excel, com a sequência de menu: `File → Open`
2. Identifique o tipo de problemas que encontra nas seguintes colunas
    a. Collector, Start Date, País, Província, Distrito, Locality Name, Collection
3. Feche o ficheiro sem gravar
4. Abra um novo Book em Excel.
5. Importe o ficheiro csv com a sequência de menu: `Data → From` text
6. Seleccione a Origem do Ficheiro, de forma a que o os acentos, cedilhas, etc, apareçam corretamente codificados. Pode ter de experimentar mais do que um tipo de origem. Os mais comuns são Unicode (UTF-8), Western Europe, Portuguese. Neste caso, trata-se de UTF-8. Para saber mais sobre a codificação de caracteres, consulte a Wikipedia[1].
7. Selecione o delimitador de coluna correcto. Os mais comuns são o tabulador, a vírgula, o ponto e vírgula. Conclua a importação e veja o ficheiro.
8. Feche o ficheiro sem gravar

Tarefa 2. Abrir um ficheiro csv em LibreOffice

1. Abrir o ficheiro `dados_sampleQLDB_ex01_code01.csv` com o LibreOffice Calc, com a sequência de menu: `File → Open`
2. Selecione o Character set - UTF-8 - e separador - vírgula - no painel de importação de texto que aparece. verificar na pré-visualização se o ficheiro foi bem interpretado. Carregar em OK.

Discussão

Quais as diferenças na abertura dos ficheiros que encontrou entre os dois programas?

Tarefa 3. Gravar tabelas de dados em formatos de texto csv

A gravação de ficheiros em formatos csv (comma separated values) é uma das melhores formas de garantir a compatibilidade com outros programas. Por vezes, é preciso também alterar a codificação de caracteres do ficheiro, para garantir compatibilidade. Neste exercício, pretende-se abrir o ficheiro `dados_sampleQLDB_ex01_code02.csv`, e gravá-lo com uma codificação UTF-8.

[1] https://pt.wikipedia.org/wiki/Codifica%C3%A7%C3%A3o_de_caracteres

1. Abrir o ficheiro `dados_sampleQLDB_ex01_code02.csv` com o Excel, com o procedimento de importação descrito acima. Deve descobrir qual o código de caracteres e separador de colunas corretos.
2. Gravar o ficheiro com "Save as", com o nome `dados_sampleQLDB_ex01_code02_utf8.csv` e escolhendo o formato CSV UTF-8 (Comma delimited) (.csv)
3. Confirme que o ficheiro foi gravado com a codificação de caracteres pretendida:
    a. Volte a importar em excel, e verifique que a codificação UTF-8 resulta na interpretação correcta dos caracteres
    b. Abra o ficheiro com o Notepad++, e verifique que na barra de informação inferior a codificação detectada é UTF-8
4. Agora, repita os passos anteriores, abrindo o ficheiro `dados_sampleQLDB_ex01_code02.csv` com o LibreOffice.
5. Faça "`Save as`", adicione "`_utf8`" ao nome, e seleccione "`Edit filter settings`" nas opções abaixo da seleção do tipo de ficheiro
6. Verifique que pode seleccionar vários formatos de codificação de caracteres, de delimitadores de campos, e de delimitadores de texto (string). Grave como UTF-8, e separado por vírgulas
7. Verifique no Notepad++ que o ficheiro gravado tem a codificação UTF-8

Discussão

Qual a importância de controlar a codificação de caracteres em conjuntos de dados?

Parte 2. Operações comuns de qualidade de dados

Tarefa 4. Identificar e limpar espaços em branco

Os espaços em branco são muitas vezes introduzidos nos dados de forma inadvertida. Para o computador, funciona como um caractere, influenciando a correspondência entre valores, ou a ordenação de listas. Os espaços em branco devem ser removidos.

1. Faça uma cópia do de ficheiro `dados_sampleQLDBP_ex01.xls` para realizar o seu exercício.
2. Abra esta cópia com a sua folha de cálculo de preferência, Excel ou Libreoffice.

3. Insira uma coluna à esquerda, com o nome `ID`, e numerada em série. Esta coluna serve para poder voltar à ordenação inicial do ficheiro.
4. Ordene por ordem, crescente a coluna `Collectors`. Assegure-se de que a área seleccionada para a ordenação inclui todas as colunas de dados. A falha de selecção de todas as colunas de dados, aquando de uma ordenação, é uma das causas comuns de erros, e grave se não verificada a tempo.
5. Verifique, percorrendo a coluna, que a ordem dos colectores é afectada pelos espaços em branco. Pode também verificar adicionando um filtro automático (Autofilter)
6. Remova os espaços em branco que existem no início e fim do texto. Para isso
    a. Adicione uma coluna em branco à direita da coluna collectors
    b. Aplique a fórmula `TRIM` (em PT, `ARRUMAR`), usando como parâmetro a célula da coluna `Collectors`.
    c. Copie esta coluna e cole sobre a coluna Collectors, com o colar especial apenas de valores.

Tarefa 5. Criar a coluna eventDate e formatar as datas com o formato de data Darwin Core, a partir da coluna StartDate

Sabendo que o formato ISO padrão de data recomendado é AAAA-MM-DD (https://dwc.tdwg.org/terms/#dwc:eventDate), mas que o a coluna contém os dados no formato DD/MM/AAAA, e que nem todas as datas estão completas, havendo casos em que apenas contém mês e ano, e outros em que contém apenas ano. São também formatos ISO padrão as seguintes opções: AAAA (quando só se indique ano) e AAAA-MM (quando se indique ano e mês)

1. Criar uma nova coluna `EventDate`, à direita da coluna `StartDate`, e copiar os valores desta última coluna
2. Seleccionar a coluna, e depois no menu `Home → Format number`, seleccionar `More Number Formats`, se estiver a usar Excel. Em Libreoffice será o menu `Format → Cells`
3. Em Excel, selecionar a categoria de formato `Custom`, e depois inserir `yyyy-mm-dd`. Em Libreoffice, pode fazê-lo no campo `Format code`, quer na categoria `Date`, como em `User-defined`.
4. Verificar o resultado, em particular para os valores em que a data não é completa, apenas contendo ano/mês, ou apenas ano.
5. Definir uma estratégia para que nestes últimos casos os registos apareçam com valores yyyy-mm ou yyyy. Note-se que a coluna StartDatePrecision pode ser usada como classificador do tipo de data.

Tarefa 6. Criar a coluna eventDate com o formato de data Darwin Core a partir da coluna StartDate - opção 1

Outra opção será obter em separado os valores do ano, mês e dia, e depois combiná-los no formato pretendido:

1. Separar os valores em três novas colunas, através do menu `Data → Text` to columns, usando o caracter "/" como separador de colunas.
2. Criar uma nova coluna chamada `EventDate`. Nesta coluna, combinar os valores de ano, mês e dia por esta ordem, no formato AAAA-MM-DD. Pode ser usada a fórmula `=AA&"-"&MM&"-"&DD`, em que AA indica a célula que contém o valor do ano, MM a célula com o valor do mês, e DD a célula com o valor do dia. Pode ser também usada a função `CONCATENATE` (em PT chama-se `CONCATENAR`). Neste caso, é necessário verificar como irá resultar a fórmula nos casos em que a data não é completa. Algumas opções para a resolução automática deste problema é usar a função condicional ISBLANK, para ignorar os casos em que os valores não existem.
3. Outra opção é usar a função `TEXTJOIN`. Neste caso, os parâmetros da função pedidos são o separador a utilizar, se os vazios são para ignorar (TRUE ou FALSE), e finalmente as colunas em que se encontram os valores a combinar. Esta função é ideal para resolver automaticamente as ausências de valores aquando da combinação de texto de várias colunas.

Tarefa 7. Usar uma lista de referência para preencher automaticamente valores

Muitas vezes é necessário completar dados relativamente a um determinado valor. Um exemplo é o preenchimento dos vários níveis taxonómicos para uma espécie, do Reino ao Género, ou dos vários níveis administrativos, do País ao município, para o caso da freguesia. Isto pode ser feito usando listas de referência e a função VLOOKUP (em PT, PROCV), que permite estabelecer correspondências entre tabelas.

1. Abra o ficheiro `lista_nomesCientificos_ex01.xls`
2. Copie toda a tabela para uma nova folha do ficheiro de dados `dados_sampleQLDBP_ex01.xls` que usou até agora. Chame a esta folha `taxon`
3. Na tabela original, insira 5 novas colunas antes da coluna Full Name, e dê-lhes os seguintes nomes: Phylum, Class, Order, Family, Genus

4. Na coluna Phylum, insira a seguinte fórmula `=VLOOKUP(AA, BB:CC, 4,0)`, em que AA corresponde à célula que contém o nome do taxon para o qual se pretende obter o valor do phylum, BB:CC à tabela que contém na primeira coluna o taxon de correspondência e nas colunas seguintes os valores que se pretendem obter, 3 é o nº de ordem da coluna que contém o valor que se pretende obter, e 0 corresponde ao valor FALSO relativamente aos valores das relação estarem ordenados. Neste caso, se a coluna Full Name, que tem os nomes dos taxa que se pretende corresponder, estiver na coluna T, então a fórmula deverá ser, para a linha 2, `=VLOOKUP(T2,$taxon.A$1:H$398,4,0)`. A utilização do símbolo $ é importante para fixar o endereço da célula de início e fim da tabela que fornece a informação.
5. Copie a fórmulas para todas as restantes linhas da coluna.
6. Repita esta operação para as colunas `Class` a `Genus`. Não esqueça de ajustar o nº de ordem da coluna da qual pretende obter o valor de cada coluna.
7. Verifique, percorrendo a tabela, que os valores ficaram bem preenchidos.

Tarefa 8. Exportar para csv e verificar a tabela

1. Gravar o ficheiro no formato csv, conforme exercício anterior, com a codificação UTF-8.
2. Abrir o ficheiro csv en Notepad++, e verificar que:
    a. a codificação dos caracteres
    b. o número de linhas do ficheiro corresponde ao nº de linhas da tabela em Excel ou Libreoffice
    c. que não existem espaços em branco no início dos valores dos colectores

Discussão

Como podem os procedimentos realizados ajudar na melhoria da qualidade de dados