ENAP Escola Nacional de Administração Pública

Programa de Aperfeiçoamento em Gestão de Políticas de Proteção e Desenvolvimento Social

Curso: Gestão da Informação

OpenRefine – Uma ferramenta grátis, de código aberto para trabalhar com dados bagunçados

A free, open source, powerful tool for working with messy data

Instrutores:

- Roberto Wagner
- Alex Pereira

Brasília, 30 de Janeiro de 2018.

Introdução ao OpenRefine

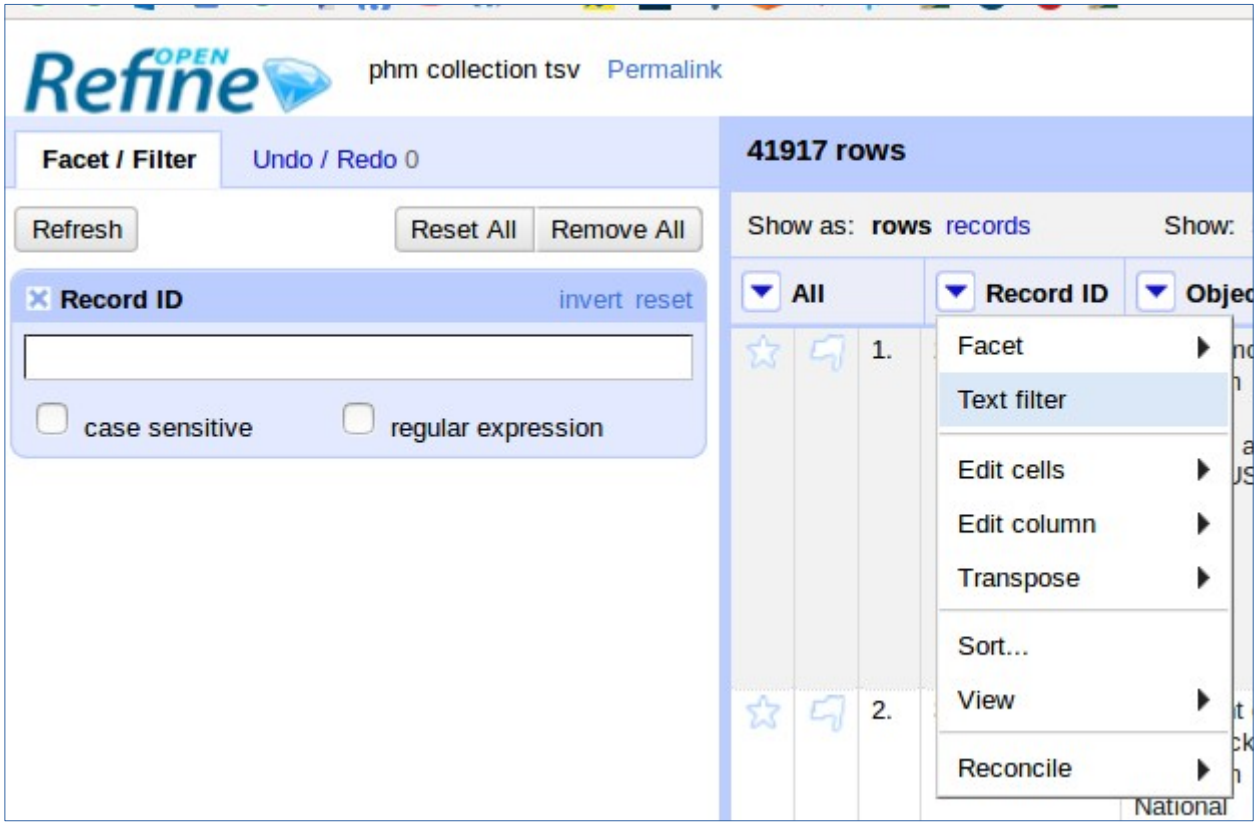
- **Ferramenta livre e Gratuita**
- **Os dados são processados localmente**
  - Não há envio de dados para nenhum serviço online/núvem
    - Dados sensíveis podem ser tratados com a ferramenta
- **A ferramenta realiza operações de ETL**
  - E – Extração (Extraction)
  - T – Transformação (Transformation)
  - L - Carregamento (Loading)

Introdução ao OpenRefine

- **Realiza “limpeza” dos dados**
  - de forma semi-automática
- **Ferramenta prática para juntar (join) dados relacionados**
  - Por exemplo, colocar numa mesma tabela:
    - População dos municípios do Brasil;
    - Quantidade de escolas de cada município; e
    - Quantidade de hospitais de cada município.

Funções básicas do OpenRefine

- **Filtro (busca)**
  - Filtra os resultados
    - Pode-se aplicar operações sobre os resultados
      - Por exemplo: remover os itens filtrados



Funções básicas do OpenRefine

- **Facets: Encontrar inconsistências**
  - Numéricas, textuais, linha do tempo, registros duplicados, entre outros...
- **Remover registros duplicados**
- **Editar células**
  - Quebrar células multi-valoradas
  - Juntar registros para formar células multi-valoradas
- **Desfazer as últimas operações**

Funções básicas do OpenRefine

- **Facets: exemplo**
  - Encontrar as diferentes categorias dos dados do Museu Powerhouse (www.powerhousemuseum.com)
  - Remover erros fazendo os dados ficarem consistentes

Referências

- https://programminghistorian.org/lessons/cleaning-data-with-openrefine
- https://programminghistorian.org/assets/phm-collection.tsv