Clean data like a pro with OpenRefine

DataHarvest 2023 - June 1-4

Anuška Delić & Hervé Letoqueux

Beginner workshop

https://s.42l.fr/DH2023_1

Dataharvest – the European Investigative Journalism Conference

OpenRefine

What is OpenRefine.

Apart from being a tool that has changed our lives...

- A powerful free, open source tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data;
- Accepts millions of rows, scriptable transformations;
- Huge fanbase(scientists, datajournalists, etc.), well documented;
- Will definitely change your life!

Important : A 90mn workshop is obviously not enough to master OpenRefine, but should be sufficient to give you a good idea of its interest, and both Anuska and I definitely want to share our love for Refine, a tool that has changed our life.

Installing OpenRefine.

It's so simple...

- Simply download OpenRefine from the official OpenRefine website or their Github repository. It works on Windows, Mac or Linux.
- OpenRefine is client/server application. It means that when you start the program, it launches a little web server in the background, and you access the application via your web browser using the URL http://127.0.0.1:3333 or http://localhost:3333.
- **Please note that a regular use of OpenRefine for its basic-yet-powerful functions doesn't need an internet connection. Your data are safely stored on your local computer.**

Let's discover OpenRefine by practicing it!

Our first Refine project...

We will use an extract from the Flickr database, a csv document : BL-Flickr-Images-Book.csv

- Create a new project :
  - **Browse** your file system, select the file, click **Next**.
  - Rename your project, add some tags.
  - Parse X rows for column headers
  - If the preview rendering is OK, click **Create project**

At this point, your dataset is IN OpenRefine as a **project** : it has 8287 rows, 15 columns, paginated 10/10 by default.

- Let's **facet** by text the dataset based on the **Place of publication** by clicking the little triangle on top of this column : **Facet->Text Facet**. On the little widdow now visible on the left, click **count**, to order by count.

Clustering data with OpenRefine.

The WOW effect...

A dataset always contains messy data : place here contains typos, mispellings, more than two cities... Let's quickly handle that!

- On the menu of the column, select **Edit cells->Cluster and Edit**
- By default, using the Fingerprint algorithm, Refine identifies matching cities and results. It also suggests a matching resulted based on the max number of occurencies.
- Where the values are correct, click the **Merge** case. Once every value is corrected, click on **Merge selected and recluster**. All these value are now corrected!
- Try other algorithms!

Clustering data is really a fantastic feature that will save you hours of pain!

Let's play with the UX of OpenRefine.

Strange one but you'll get used to it...

- Menu of columns
- The **All** column.
- The **Export** menu, to export a dataset or a project (a dataset with all the transfomations you made!)
- The **Common transforms** menu : sanitizing spaces, Capitalize, etc...
- The **Sort** menu (temporary and permanent)
- The **Filter** menu (text, regexp). Extract a subset.
- How to **collapse/uncollapse** columns
- Undo/Redo

Powerful functions seem sometimes hidden or difficult to find. But every new versions of OpenRefine sees its UX enhanced!

Let's dive into some cool transformations (I).

We are pretty sure you already like it...

We want to create a new column named **bracketed**, that will contains every value of **Date of Publication** that has a bracket sign in it.

- In the column menu, select **Text filter**. In the little windows on the left select regular expression and type \[.*\] (it means : any caracters between two brackets). You should get 995 matching rows.
- In the column menu, click on **Edit column->create column based on this column**. Give it a name (**bracketed**), and click **ok**.
- Remove the regexp filter by closing the little window on the left.

Always remember that you transformations will only apply on the selected part of the dataset (here, the filtered one). So always double-check the active rows!
But for now, let's **Undo** these last steps and get back to the original dataset.

Let's dive into some cool transformations (II).

Where we learn our first GREL command...

GREL is a simple language to perform transformations on your data. 6 or 7 command will help you in 90% of your needs; don't panic! The most useful one is **value.partition()** that breaks a string in 2 parts around a pivot. Explaination :

- Suppose you have the email *john.doe@gmail.com*. It stands for *nickname AT domain*. If I want to retrieve the domain, I can split the string around the @ sign (our pivot) and take everything that is on the right of the pivot.
- In GREL, this is written like this : **value.partition("@")[2]** : the pivot is into quotes. [0] means left, [2] means right. [1] is the pivot.
- To sum up : we take the original value (**value**), we add the GREL command **partition** with a dot.

Useful commands : slice(), split(), join(), find()...
See the Help menu to get the syntax.

Let's dive into some cool transformations (II).

Where we learn our first GREL command...

Let's use this to create a new column based on the **Date of publication**, containing only the dates without brackets.

- Click on the menu of the column: **Edit column->add column based on this column**, give it a name
- In the value section, type : **value.partition(" [")[0]** or **value.partition(/\[.*\]/)[0]**. The two slashes are here to say to openRefine that it's a regexp and not an exact expression.
- The preview should give you an idea of what is happening here. Click **[**Ok] when done.

There are always several ways of doing transformations in OpenRefine!

Reconciliation.

How to enrich and consolidate your datasets...

OpenRefines Reconciliation service is used to semi-automate the process of matching data in OpenRefine fields with more authoritative data in external sources.
In order ro reconcile our data, we need an access to an external reconciliation service. There are plenty of them on the Internet!

- The OpenSanctions reconciliation service
  `https://api.opensanctions.org/reconcile/default`
- Geonames `https://fornpunkt.se/apis/reconciliation/geonames`
- A complete list of services...
  `https://reconciliation-api.github.io/testbench/`

Reconciliation.

How to enrich and consolidate your datasets...

We will use the simple dataset named PRESIDENTS.CSV, a list of the last 10 presidents of the USA. We want to retrieve some consolidated data on them such as their place of birth and death. Wikipedia knows that.

- Create a project by using the PRESIDENTS.CSV file.
- On column 2, menu, click on **Reconcile->start reconciling**, select the wikidata service. Select Human, auto-match, and maximum number of candidates : 2. Click **Start reconciling**.
- Some presidents are quickly recognized while some are not. Select the good entry by a rollover ith the mouse.
- Now all the presidents should be"in blue", as links.
- On the menu of the column, select **Edit columns->creating new columns from reconciled values.** and choose the data you want to import!

Let's wrap up!

We have learnt a lot!!!

You are now more familiar with OpenRefine!

- We've learnt how to quickly cluster and cleanse a dataset;
- We discovered a part of the UX of OpenRefine
- We have dived into the transformations world, by simple clicks or by using one command of the GREL language
- We have enriched our dataset with consolidated data by using reliable external sources.

Congratulations, that's HUGE!!!

One last thing...

Discover the Refine recipes

With openRefine, every step, transformation you make can be easily repeated. It's called a **recipe**.

- Use the "presidents" project. Click on **Undo/Redo**, and click on **Extract**.
- You can see here all the steps you made. Select everything on the left, copy/paste it in a text file on you computer. This is a recipe that you can now use on all similar datasets!
- create a new project, reimport the "presidents.csv" file, Click on **Undo/Redo**, and click on **Apply**
- Paste your recipe and click OK.

OpenRefine is magic!!!! \o/

Questions

OpenRefine, je res super!
OpenRefine, c'est vraiment génial!