

# Le nettoyage de données!

# Clean your Data!

Constance Crompton

Chaire de recherche du Canada en humanités numériques

Canada Research Chair in Digital Humanities

Diapo/ Slides CC BY-NC-SA





# OpenRefine

## Dans ces diapos

- OpenRefine: Instructions d'installation
- OpenRefine: des instructions de base
- OpenRefine: instructions avancées (APIs)
- Ressources supplémentaires
- Merci!

## In this deck

- OpenRefine installation instructions
- OpenRefine basics
- OpenRefine: going further (APIs)
- Further Resources
- Thank you!

# OpenRefine



**OpenRefine est une application de nettoyage de données qui s'exécute dans votre logiciel de navigation (Firefox ou Chrome).**

Instructions d'installation et vidéos tutorielles :  
<https://openrefine.org/> (choisissez la version applicable - mac, win, ou linux)

Suivez les instructions d'installation

**Pour ouvrir OpenRefine**, double-cliquez sur l'application à l'endroit où votre installation l'a enregistrée (probablement dans Applications)

Pour fermer OpenRefine dans Windows : Control-C

Pour fermer OpenRefine sur un Mac : Cliquez sur l'application OpenRefine dans la dock, puis appelez Quit

**OpenRefine is an data cleaning application that runs in your browser (Firefox or Chrome).**

Download and install OpenRefine.

Installation instructions and tutorial videos:  
<https://openrefine.org/> (choose mac or win versions (or linux if that is what you are running))

Follow the installation instructions

**To open OpenRefine**, double click on the application wherever your installation saved it (likely in Applications)

To shut down OpenRefine in Windows: Control-C

To shut down OpenRefine on a Mac: Click the OpenRefine app in the dock, invoke Quit



# OpenRefine: des instructions de base/basics

Télécharger

[https://www.dropbox.com/s/ysz6f3th0eojvbo/concerts\\_cleaning\\_nettoyage.csv?dl=0](https://www.dropbox.com/s/ysz6f3th0eojvbo/concerts_cleaning_nettoyage.csv?dl=0)

Download

[https://www.dropbox.com/s/ysz6f3th0eojvbo/concerts\\_cleaning\\_nettoyage.csv?dl=0](https://www.dropbox.com/s/ysz6f3th0eojvbo/concerts_cleaning_nettoyage.csv?dl=0)

21538 rows Extensions: **RDF** Wikidata

Show as: **rows** records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

All	Date	Location	Time	Venue	eventType	season	programID	orchestra	id
☆	1.	1842-12-07T05:00:00Z	Manhattan, NY	8:00PM	Apollo Rooms	Subscription Season	1842-43	3853	38e072a7-8fc9-4f9a-8eac-3957905c0002
☆	2.	1843-02-18T05:00:00Z	Manhattan, NY	8:00PM	Apollo Rooms	Subscription Season	1842-43	5178	c7b2b95c-5e0b-431c-816-51231-2001-1
☆	3.	1843-04-07T05:00:00Z	Manhattan, NY	8:00PM	Apollo Rooms	Special	1842-43	10785	7-aec0-
☆	4.	1843-04-22T05:00:00Z	Manhattan, NY	8:00PM	Apollo Rooms	Subscription Season	1842-43	5887	6-9831-
☆	5.	1843-11-18T05:00:00Z	Manhattan, NY	None	Apollo Rooms	Subscription Season	1843-44	305	6-
☆	6.	1844-01-13T05:00:00Z	Manhattan, NY	8:00PM	Apollo Rooms	Subscription Season	1843-44	3368	9
☆	7.	1844-03-16T05:00:00Z	Manhattan, NY	None	Apollo Rooms	Subscription Season	1843-44	4226	3a-ae2a-
☆	8.	1844-05-18T05:00:00Z	Manhattan, NY	None	Apollo Rooms	Subscription Season	1843-44	5087	a-b8d2-
☆	9.	1844-11-16T05:00:00Z	Manhattan, NY	8:00PM	Apollo Rooms	Subscription Season	1844-45	6310	a6c4e2612c
☆	10.	1845-01-11T05:00:00Z	Manhattan, NY	8:00PM	Apollo Rooms	Subscription Season	1844-45	1979	b- 03aa4112ze659e56 253d22e1-9d44-410c-ae06-61abe434e5ec

Context menu for 'orchestra' column:

- Facet
- Text filter
- Edit cells
  - Transform...
  - Common transforms
    - Fill down
    - Blank down
    - Split multi-valued cells...
    - Join multi-valued cells...
  - Cluster and edit...
- Edit column
- Transpose
- Sort...
- View
- Reconcile
- Replace

Column Arrow > Edit Cells > Cluster and Edit

# OpenRefine



## Cluster & Edit column "orchestra"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method

Keying Function

2 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	12	<ul style="list-style-type: none"><li>N. Y. Symphony (8 rows)</li><li>n. y. Symphony (4 rows)</li></ul>	<input checked="" type="checkbox"/>	<input type="text" value="N. Y. Symphony"/>
2	15	<ul style="list-style-type: none"><li>New York Symph. (8 rows)</li><li>New York Symph (7 rows)</li></ul>	<input checked="" type="checkbox"/>	<input type="text" value="New York Symph."/>

# Rows in Cluster



12 — 15

Average Length of Choices



14 — 14.5

Length Variance of Choices



0 — 0.5

Column Arrow > Edit Cells > Cluster and Edit

# OpenRefine



*Column Arrow > Edit  
Column > Split into  
several columns*

... et après renommer la  
colonne  
... and then rename the  
column

21538 rows

Show as: rows records Show: 5 10 25 50 rows

All	Date	Location	Time	Venue	eventType
1.	Facet	Manhattan, NY	8:00PM	Apollo Rooms	Subscription Season
2.	Text filter	Manhattan, NY	8:00PM	Apollo Rooms	Subscription Season
3.	Edit cells	Manhattan...	8:00PM	Apollo	Social
4.	Edit column	Split into several columns...			ription Season
5.	Transpose	Join columns...			ription Season
6.	Sort...	Add column based on this column...			ription Season
7.	View	Add column by fetching URLs...			ription Season
8.	Reconcile	Add columns from reconciled values			
8.	1844-05-18T05:00:00Z	Rename this column			
9.	1844-11-16T05:00:00Z	Remove this column			
10.	1845-01-11T05:00:00Z	Move column to beginning			
		Move column to end			
		Move column left			
		Move column right			

### Split column Date into several columns

**How to Split Column**

by separator  
Separator   regular expression

Split into  columns at most (leave blank for no limit)

by field lengths

List of integers separated by commas, e.g., 5, 7, 15

**After Splitting**

Guess cell type  
 Remove this column

OK Cancel



# OpenRefine avec APIs / with APIs

Wikidata: <https://www.wikidata.org>

Column Arrow (Time) > Reconcile >  
Start Reconciling

<https://wikidata.reconci.link/en/api>

Time of day

Reconcile column "Location"

Services » Access Service API

Wikidata (en)  one of these types:

VIAF  district

Wikidata reconci.link (en)

Also use relevant details from other columns:

Column	Include?	As Property
Date	<input type="checkbox"/>	<input type="text"/>
Year	<input type="checkbox"/>	<input type="text"/>
Time	<input type="checkbox"/>	<input type="text"/>
Venue	<input type="checkbox"/>	<input type="text"/>
eventType	<input type="checkbox"/>	<input type="text"/>
season	<input type="checkbox"/>	<input type="text"/>
programID	<input type="checkbox"/>	<input type="text"/>
orchestra	<input type="checkbox"/>	<input type="text"/>
id	<input type="checkbox"/>	<input type="text"/>

Add Standard Service... Start Reconciling Cancel



# OpenRefine avec APIs / with APIs

Wikidata: <https://www.wikidata.org>

Column Arrow (Time) > Reconcile >  
Start Reconciling

<https://wikidata.reconci.link/en/api>

Time of day

Reconcile column "Location"

Services » Access Service API

Wikidata (en)  one of these types: Also use relevant details from other columns:

VIAF  district

Wikidata reconci.link (en)


Column	Include?	As Property
Date	<input type="checkbox"/>	<input type="text"/>
Year	<input type="checkbox"/>	<input type="text"/>
Time	<input type="checkbox"/>	<input type="text"/>
Venue	<input type="checkbox"/>	<input type="text"/>
eventType	<input type="checkbox"/>	<input type="text"/>
season	<input type="checkbox"/>	<input type="text"/>
programID	<input type="checkbox"/>	<input type="text"/>
orchestra	<input type="checkbox"/>	<input type="text"/>
id	<input type="checkbox"/>	<input type="text"/>

1.	1842-12-07T05:00:00Z	1842	Manhattan, NY	8:00PM	edit	Apollo	Subscription	1842-43	3853	New
				<input checked="" type="checkbox"/> 20:00 (62) <input checked="" type="checkbox"/> evening (25) <input checked="" type="checkbox"/> Create new Search for match						
2.	1843-02-18T05:00:00Z	1843	Manhattan, NY	8:00PM						
				<input checked="" type="checkbox"/> 20:00 (62) <input checked="" type="checkbox"/> evening (25) <input checked="" type="checkbox"/> Create new Search for match						
3.	1843-04-07T05:00:00Z	1843	Manhattan,	8:00PM		Apollo	Special	1842-43	10785	Mus

Match this Cell  Match All Identical Cells  Cancel

20:00 (Q55812694)

point of time during the day, 08:00 pm local time, in the evening







# OpenRefine avec APIs / with APIs

Télécharger

<https://www.dropbox.com/s/y2x4mrqe1a5fu47/craikpeople.xlsx?dl=0>

Let's populate a whole spreadsheet with Wikidata

*Column Arrow > Edit Columns > Add Columns from reconciled values*

*GREL VIAF ID Column Arrow > Edit Cells > Transform*

Download

<https://www.dropbox.com/s/y2x4mrqe1a5fu47/craikpeople.xlsx?dl=0>

Remplissons une feuille de calcul entière avec des Wikidata

*Column Arrow > Edit Columns > Add Columns from reconciled values*

*GREL VIAF ID Column Arrow > Edit Cells > Transform*




# OpenRefine avec APIs / with APIs

Library of Congress API instructions

<https://libraryofcongress.github.io/data-exploration/requests.html>

Notre API call: <https://www.loc.gov/pictures/?q=canada&fo=json>

 **OpenRefine** *A power tool for working with messy data.*

Create Project « Start Over Configure Parsing Options Project name temp Tags Create Project »



Open Project  
Import Project  
Language Settings

	collections - code	collections - link	collections - thumb	collections - title	collections - thumb_large	collections - thumb_feature
1.	ahii	<a href="https://www.loc.gov/pictures/collection/ahii/">https://www.loc.gov/pictures/collection/ahii/</a>	<a href="https://www.loc.gov/pictures/static/data/ahii/thumb.png">https://www.loc.gov/pictures/static/data/ahii/thumb.png</a>	Abdul Hamid II Collection	<a href="https://www.loc.gov/pictures/static/data/ahii/thumb_large.png">https://www.loc.gov/pictures/static/data/ahii/thumb_large.png</a>	<a href="https://www.loc.gov/pictures/static/data/ahii/featured.png">https://www.loc.gov/pictures/static/data/ahii/featured.png</a>
2.	anedub	<a href="https://www.loc.gov/pictures/collection/anedub/">https://www.loc.gov/pictures/collection/anedub/</a>	<a href="https://www.loc.gov/pictures/static/data/anedub/thumb.png">https://www.loc.gov/pictures/static/data/anedub/thumb.png</a>	African American Photographs Assembled for 1900 Paris Exposition	<a href="https://www.loc.gov/pictures/static/data/anedub/thumb_large.png">https://www.loc.gov/pictures/static/data/anedub/thumb_large.png</a>	<a href="https://www.loc.gov/pictures/static/data/anedub/featured.png">https://www.loc.gov/pictures/static/data/anedub/featured.png</a>

0 row(s) of data

- Line-based text files
- CSV / TSV / separator-based files
- Fixed-width field text files
- PC-Axis text files
- JSON files**
- MARC files

- Load at most
- Preserve empty strings
- Trim leading & trailing whitespace from strings
- Parse cell text into numbers, dates, ...
- Store file source (file names, URLs) in each row





# OpenRefine exporter/ export



# Ressources supplémentaires | Further Resources

Data Carpentry EN: <https://datacarpentry.org/2015-09-21-Genentech/lessons/02-starting-with-OpenRefine.html>

Amadine Oricheta FR: <https://datahist.hypotheses.org/200>

John Little EN: <https://libjohn.github.io/openrefine/>

Miriam Posner EN:

<http://miriamposner.com/classes/dh101f17/tutorials-guides/data-manipulation/get-started-with-openrefine>

Programming Historian (Seth Van Hooland FR):

<https://programminghistorian.org/fr/lecons/nettoyer-ses-donnees-avec-openrefine>

Programming Historian (Evan Peter Williamson EN):

<https://programminghistorian.org/en/lessons/fetch-and-parse-data-with-openrefine>

**Merci - restez en contact**  
**Thank you - be in touch**

**constance.crompton@uottawa.ca | @clkcrompton**

À l'atelier prochain #Boîte à outils en SHN: <https://dhsite.org/ateliers-boite-a-outils-numeriques/>  
See you at the next #DHToolbox: <https://dhsite.org/workshops-digital-humanities-toolbox/>

