Which data?

THE MET

MoMA

What is data Refining, Wrangling, Engineering…

WHAT

- Removing duplicates
- Typos
- Joining field
- Data in wrong field
- Change of format (date)
- Encoding errors
- Join data sets
- Transpose row/column
- Enrichment

AIMS

- Integrating data
- Use & Re-use
- Sharing
- Analytics

LibreOffice
The Document Foundation

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

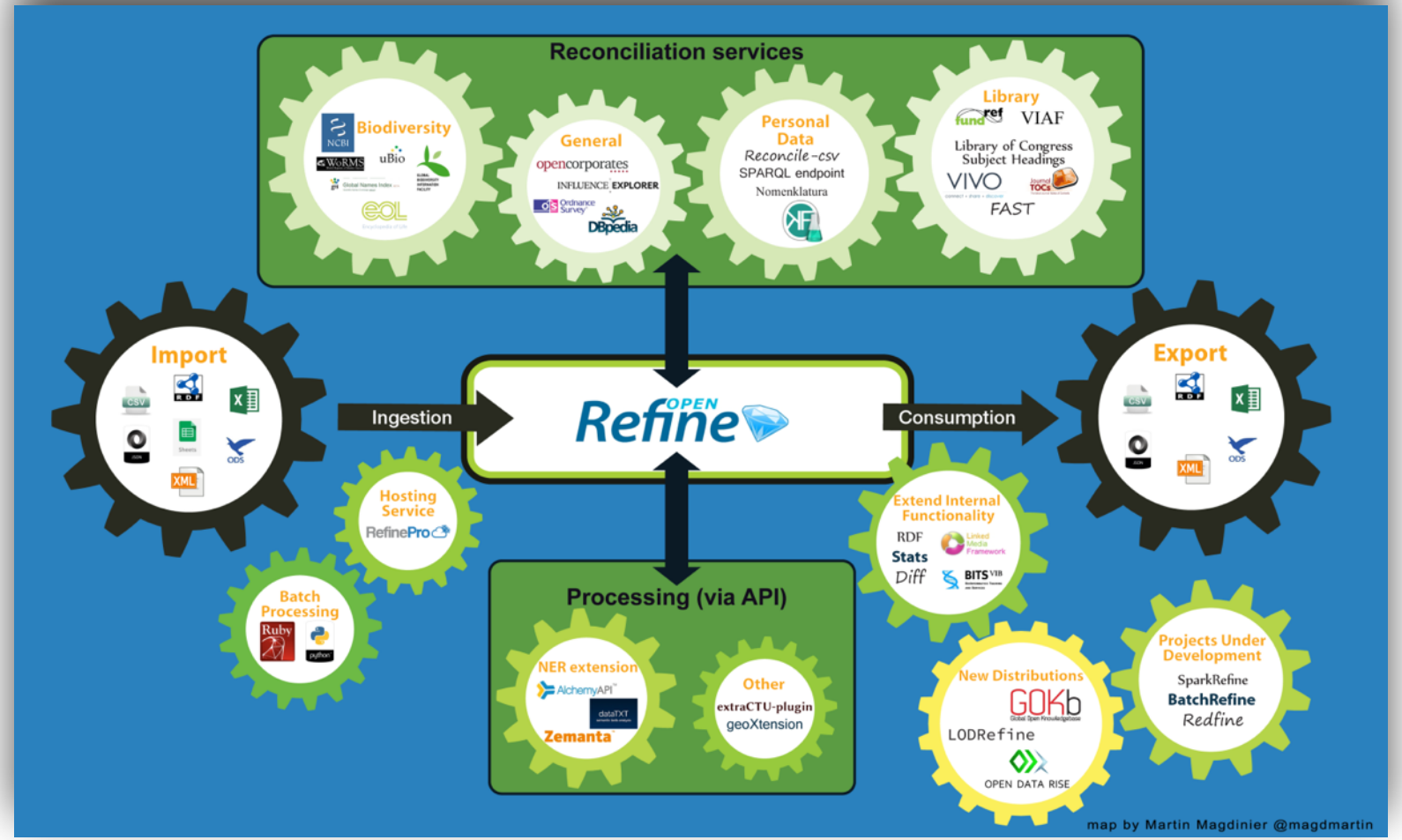
CSVKit

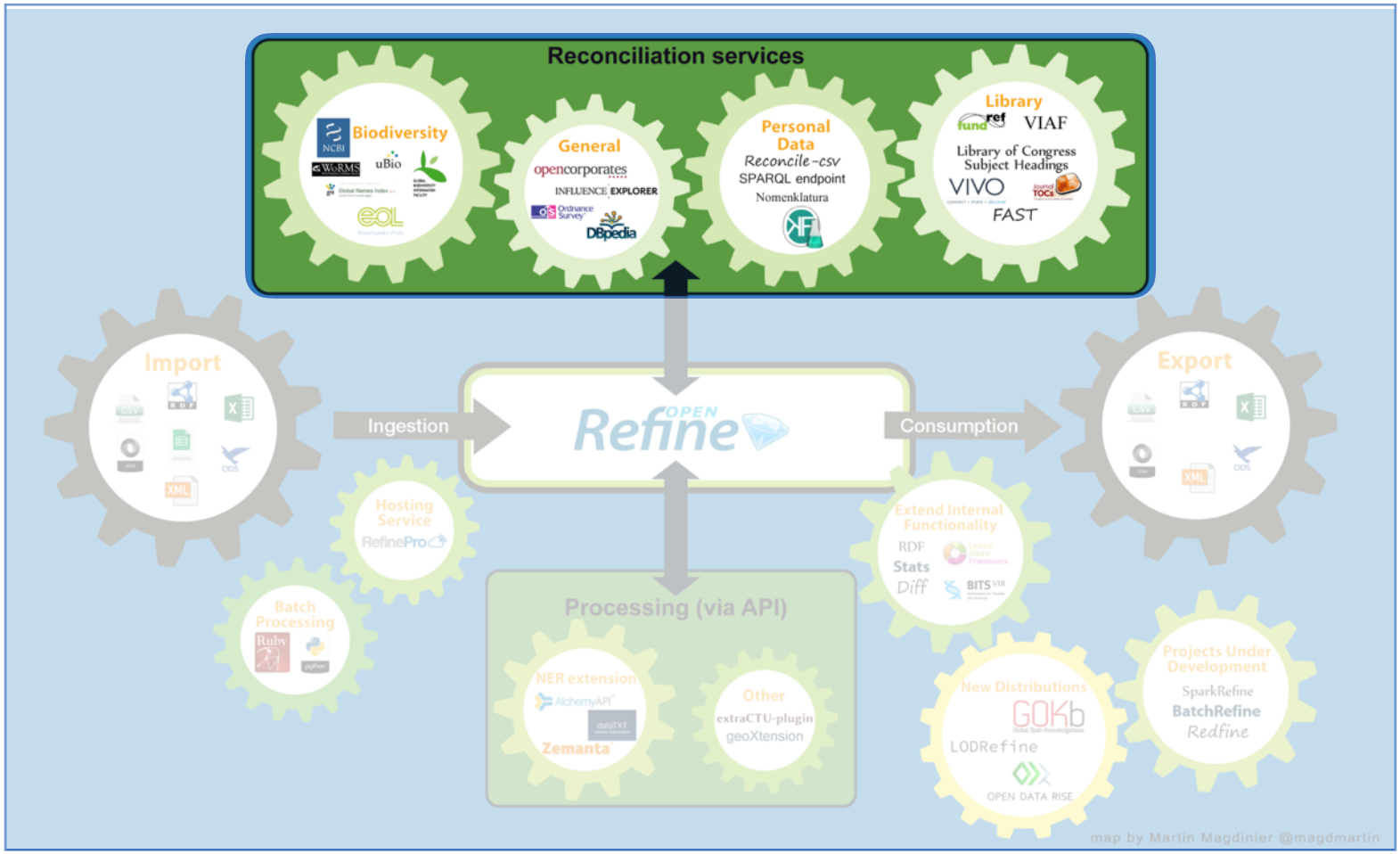
Silk

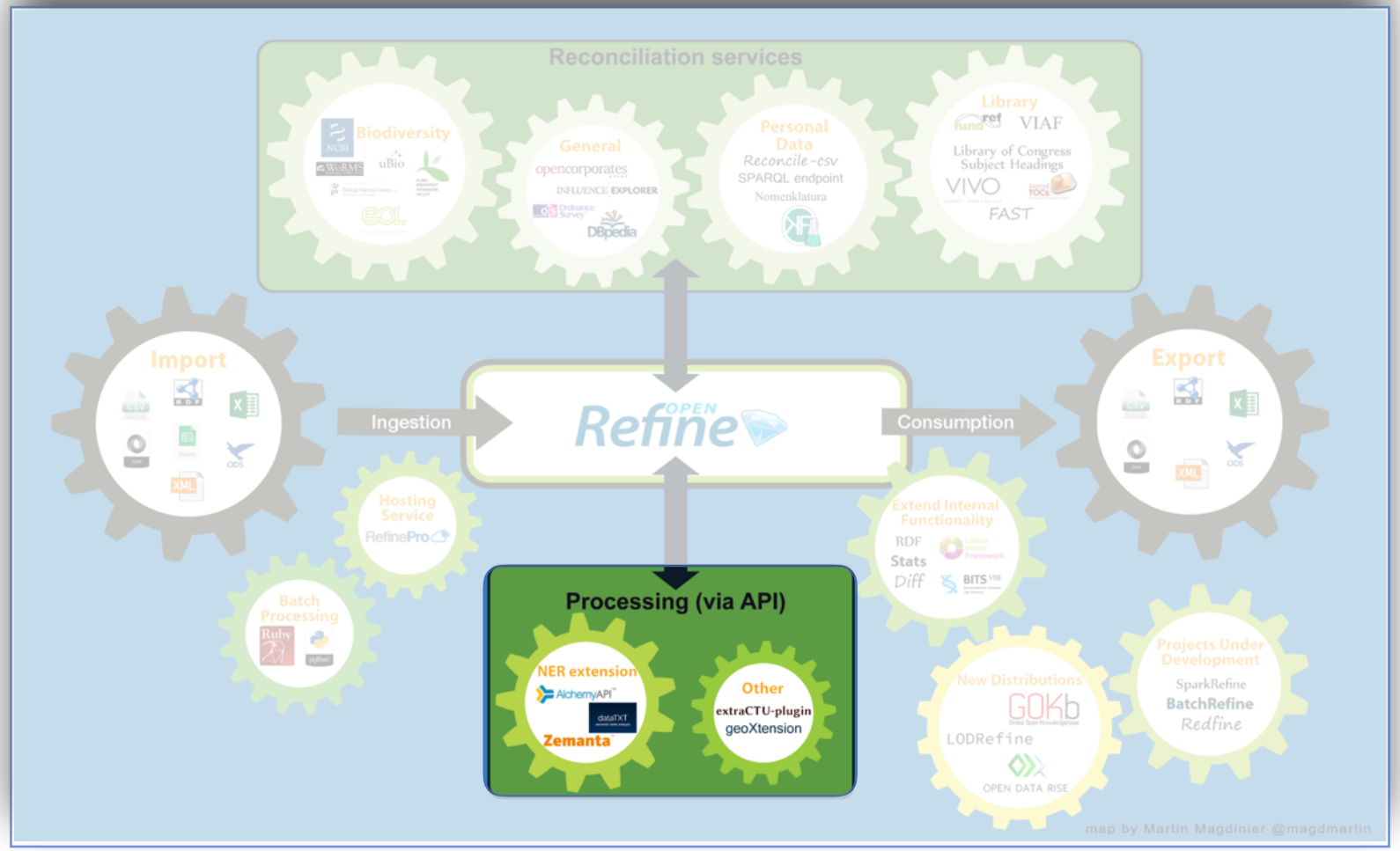
tripleGEO

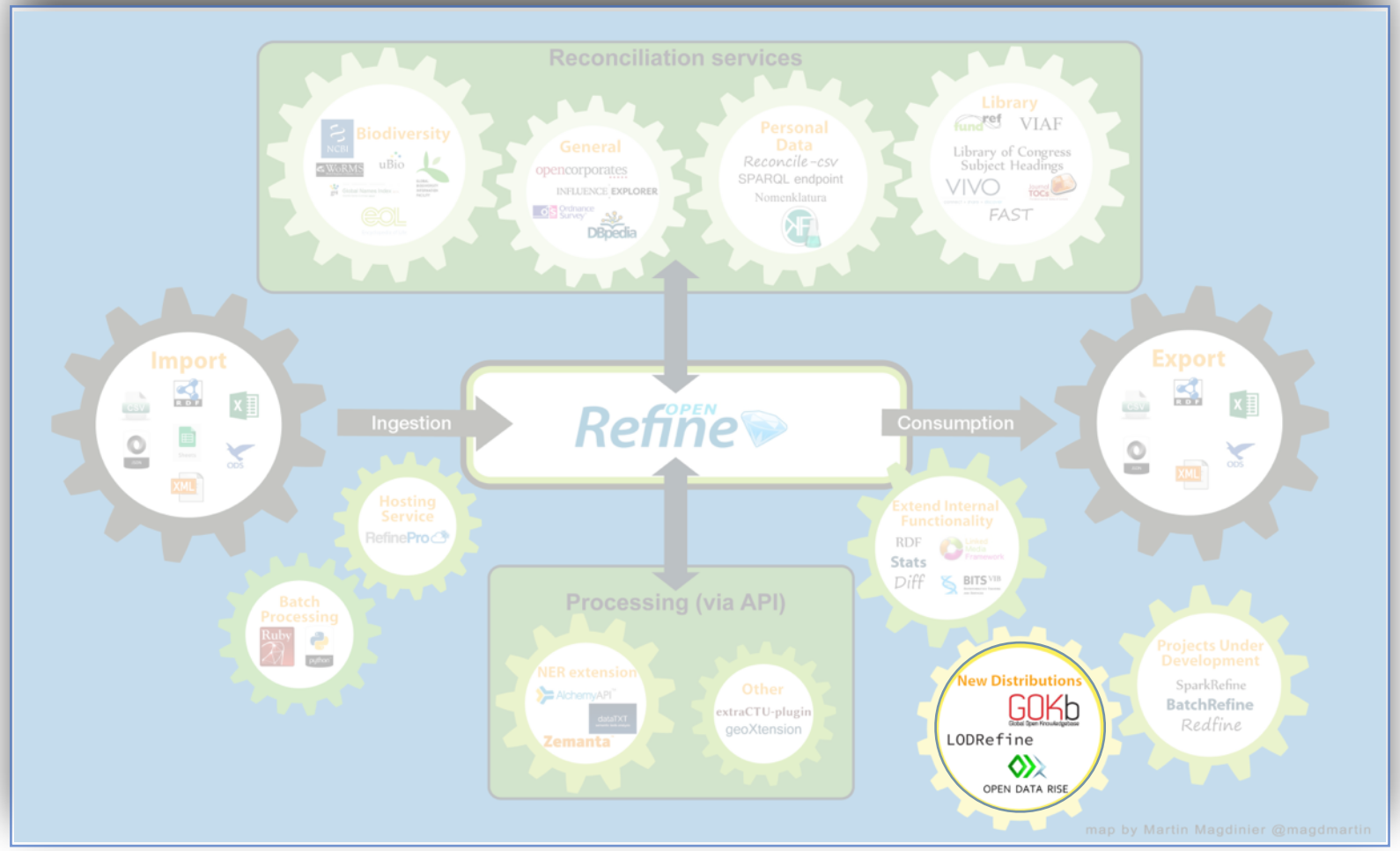
OpenRefine

- Created by MetaWeb Technology as Gridworks in 2009
- Acquired by Google and rebranded as Google Refine
- Google continue its development: version 2.0 and 2.5
- Stop supporting in 2012
- OpenRefine is born: 2013


Reconciliation services
Biodiversity
NCBI
WoRMS
uBio
eol
General
opencorporates
INFLUENCE EXPLORER
Ordnance Survey
DBpedia
Personal Data
Reconcile-csv
SPARQL endpoint
Nomenklatura
Library
fundref
VIAF
Library of Congress Subject Headings
VIVO
FAST
Import
CSV
RDF
XML
ODS
Ingestion
OPEN Refine
Consumption
Export
Hosting Service
RefinePro
Batch Processing
Ruby
python
Extend Internal Functionality
RDF
Stats
Diff
Linked Media Framework
BITS VIB
Processing (via API)
NER extension
AlchemyAPI
dataTXT
Zemanta
Other
extraCTU-plugin
geoXtension
New Distributions
GOKb
LODRefine
OPEN DATA RISE
Projects Under Development
SparkRefine
BatchRefine
Redfine
map by Martin Magdinier @magdmartin


Reconciliation services
Biodiversity
NCBI
uBio
WoRMS
eol
General
opencorporates
INFLUENCE EXPLORER
Ordnance Survey
DBpedia
Personal Data
Reconcile-csv
SPARQL endpoint
Nomenklatura
Library
fundref
VIAF
Library of Congress Subject Headings
VIVO
FAST
Import
Ingestion
OpenRefine
Consumption
Export
Hosting Service
RefinePro
Batch Processing
Ruby
python
Processing (via API)
NER extension
AlchemyAPI
dataTXT
Zemanta
Other
extraCTU-plugin
geoXtension
Extend Internal Functionality
RDF
Stats
Diff
BITS
New Distributions
GOKb
LODRefine
OPEN DATA RISE
Projects Under Development
SparkRefine
BatchRefine
Redfine
map by Martin Magdinier @magdmartin


Reconciliation services
Biodiversity
General
Personal Data
Reconcile-csv
SPARQL endpoint
Nomenklatura
Library
VIAF
Library of Congress Subject Headings
VIVO
FAST
Import
Ingestion
OPEN Refine
Consumption
Export
Hosting Service
RefinePro
Batch Processing
Ruby
Extend Internal Functionality
RDF
Stats
Diff
Processing (via API)
NER extension
AlchemyAPI
dataTXT
Zemanta
Other
extraCTU-plugin
geoXtension
New Distributions
GOKb
LODRefine
OPEN DATA RISE
Projects Under Development
SparkRefine
BatchRefine
Redfine
map by Martin Magdinier @magdmartin


Reconciliation services
Biodiversity
General
Personal Data
Reconcile-csv
SPARQL endpoint
Nomenklatura
Library
VIAF
Library of Congress Subject Headings
VIVO
FAST
Import
Ingestion
OpenRefine
Consumption
Export
Hosting Service
RefinePro
Batch Processing
Ruby
Extend Internal Functionality
RDF
Stats
Diff
Processing (via API)
NER extension
AlchemyAPI
Zemanta
Other
extraCTU-plugin
geoXtension
New Distributions
GOKb
LODRefine
OPEN DATA RISE
Projects Under Development
SparkRefine
BatchRefine
Redfine
map by Martin Magdinier @magdmartin

Data: where?

- Datahub http://datahub.io
- Europeana http://www.europeana.eu
- Github https://github.com
- EU OpenData portal https://www.europeandataportal.eu
- Scraping

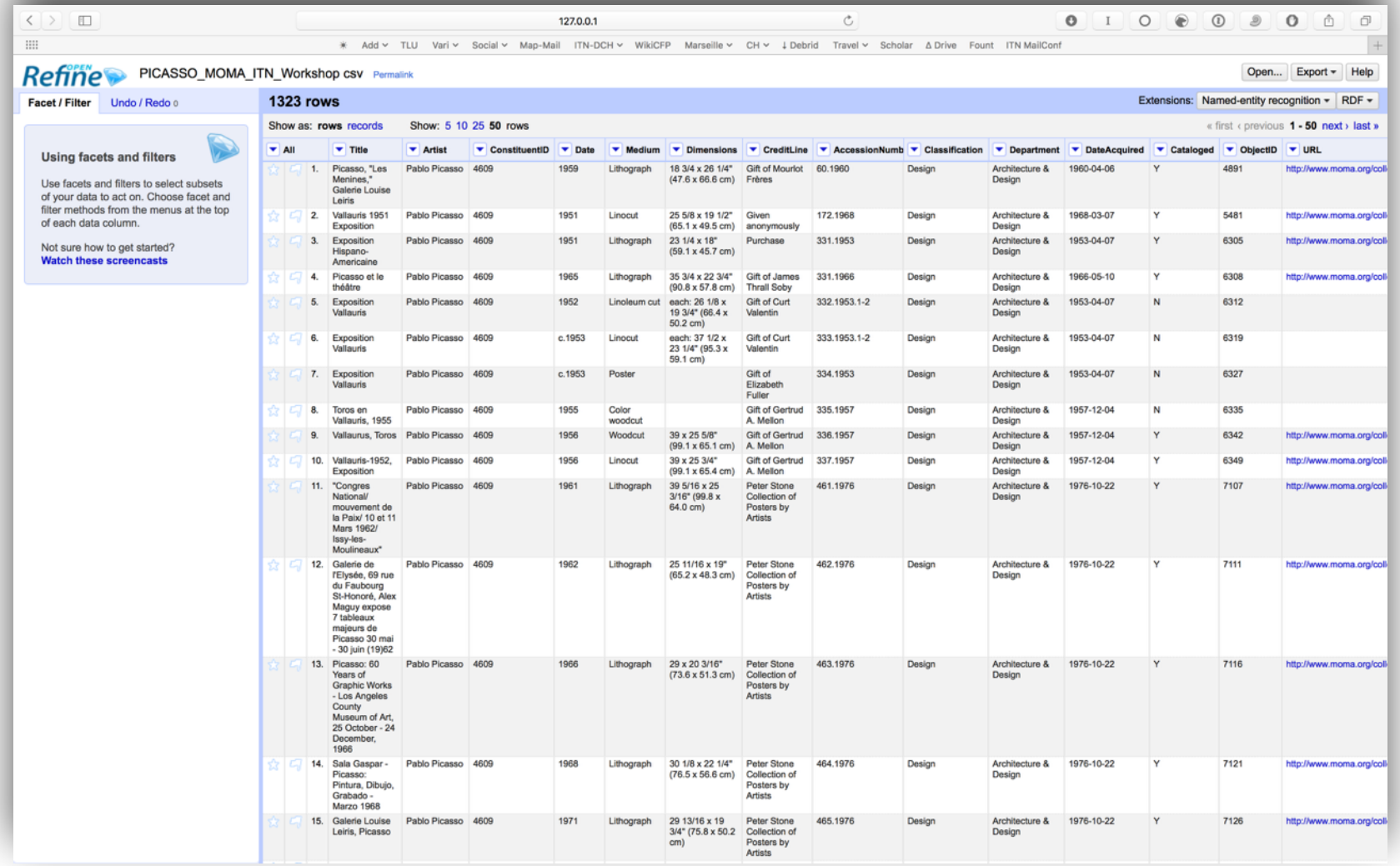

THE
MET

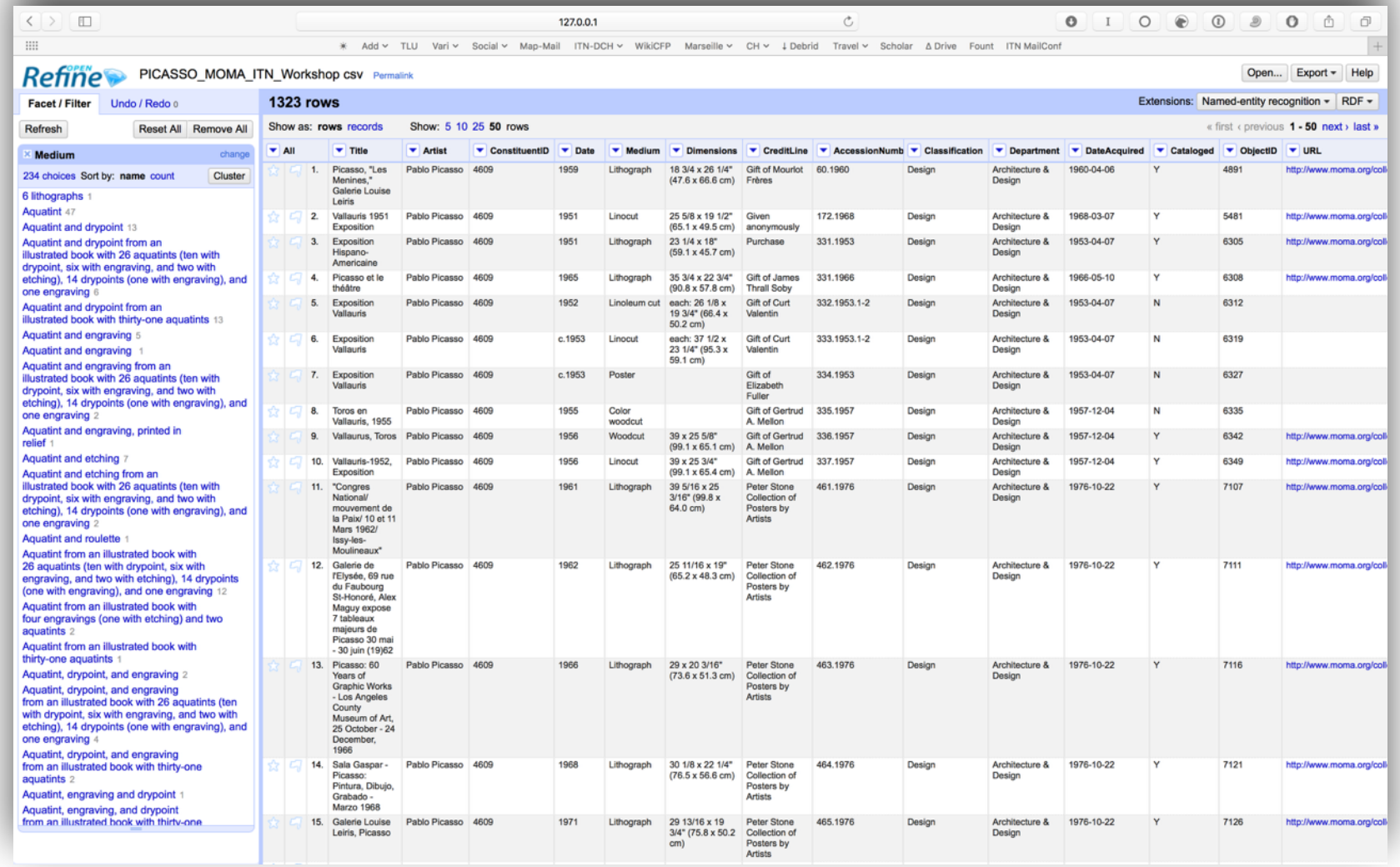

MoMA

Install Vagrant Virtual Machine

https://github.com/ncarboni/ITNDCH_workshop

127.0.0.1:3333


127.0.0.1
Add TLU Vari Social Map-Mail ITN-DCH WikiCFP Marseille CH Debrid Travel Scholar Drive Fount ITN MailConf
OpenRefine
PICASSO_MOMA_ITN_Workshop csv
Permalink
Open... Export Help
Facet / Filter
Undo / Redo 0
Using facets and filters
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.
Not sure how to get started?
Watch these screencasts
1323 rows
Extensions: Named-entity recognition RDF
Show as: rows records
Show: 5 10 25 50 rows
« first ‹ previous 1 - 50 next › last »
All
Title
Artist
ConstituentID
Date
Medium
Dimensions
CreditLine
AccessionNumb
Classification
Department
DateAcquired
Cataloged
ObjectID
URL
1. Picasso, "Les Menines," Galerie Louise Leiris | Pablo Picasso | 4609 | 1959 | Lithograph | 18 3/4 x 26 1/4" (47.6 x 66.6 cm) | Gift of Mourlot Frères | 60.1960 | Design | Architecture & Design | 1960-04-06 | Y | 4891 | http://www.moma.org/coll
2. Vallauris 1951 Exposition | Pablo Picasso | 4609 | 1951 | Linocut | 25 5/8 x 19 1/2" (65.1 x 49.5 cm) | Given anonymously | 172.1968 | Design | Architecture & Design | 1968-03-07 | Y | 5481 | http://www.moma.org/coll
3. Exposition Hispano-Americaine | Pablo Picasso | 4609 | 1951 | Lithograph | 23 1/4 x 18" (59.1 x 45.7 cm) | Purchase | 331.1953 | Design | Architecture & Design | 1953-04-07 | Y | 6305 | http://www.moma.org/coll
4. Picasso et le théâtre | Pablo Picasso | 4609 | 1965 | Lithograph | 35 3/4 x 22 3/4" (90.8 x 57.8 cm) | Gift of James Thrall Soby | 331.1966 | Design | Architecture & Design | 1966-05-10 | Y | 6308 | http://www.moma.org/coll
5. Exposition Vallauris | Pablo Picasso | 4609 | 1952 | Linoleum cut | each: 26 1/8 x 19 3/4" (66.4 x 50.2 cm) | Gift of Curt Valentin | 332.1953.1-2 | Design | Architecture & Design | 1953-04-07 | N | 6312 |
6. Exposition Vallauris | Pablo Picasso | 4609 | c.1953 | Linocut | each: 37 1/2 x 23 1/4" (95.3 x 59.1 cm) | Gift of Curt Valentin | 333.1953.1-2 | Design | Architecture & Design | 1953-04-07 | N | 6319 |
7. Exposition Vallauris | Pablo Picasso | 4609 | c.1953 | Poster | | Gift of Elizabeth Fuller | 334.1953 | Design | Architecture & Design | 1953-04-07 | N | 6327 |
8. Toros en Vallauris, 1955 | Pablo Picasso | 4609 | 1955 | Color woodcut | | Gift of Gertrud A. Mellon | 335.1957 | Design | Architecture & Design | 1957-12-04 | N | 6335 |
9. Vallaurus, Toros | Pablo Picasso | 4609 | 1956 | Woodcut | 39 x 25 5/8" (99.1 x 65.1 cm) | Gift of Gertrud A. Mellon | 336.1957 | Design | Architecture & Design | 1957-12-04 | Y | 6342 | http://www.moma.org/coll
10. Vallauris-1952, Exposition | Pablo Picasso | 4609 | 1956 | Linocut | 39 x 25 3/4" (99.1 x 65.4 cm) | Gift of Gertrud A. Mellon | 337.1957 | Design | Architecture & Design | 1957-12-04 | Y | 6349 | http://www.moma.org/coll
11. "Congres National/ mouvement de la Paix/ 10 et 11 Mars 1962/ Issy-les-Moulineaux" | Pablo Picasso | 4609 | 1961 | Lithograph | 39 5/16 x 25 3/16" (99.8 x 64.0 cm) | Peter Stone Collection of Posters by Artists | 461.1976 | Design | Architecture & Design | 1976-10-22 | Y | 7107 | http://www.moma.org/coll
12. Galerie de l'Elysée, 69 rue du Faubourg St-Honoré, Alex Maguy expose 7 tableaux majeurs de Picasso 30 mai - 30 juin (19)62 | Pablo Picasso | 4609 | 1962 | Lithograph | 25 11/16 x 19" (65.2 x 48.3 cm) | Peter Stone Collection of Posters by Artists | 462.1976 | Design | Architecture & Design | 1976-10-22 | Y | 7111 | http://www.moma.org/coll
13. Picasso: 60 Years of Graphic Works - Los Angeles County Museum of Art, 25 October - 24 December, 1966 | Pablo Picasso | 4609 | 1966 | Lithograph | 29 x 20 3/16" (73.6 x 51.3 cm) | Peter Stone Collection of Posters by Artists | 463.1976 | Design | Architecture & Design | 1976-10-22 | Y | 7116 | http://www.moma.org/coll
14. Sala Gaspar - Picasso: Pintura, Dibujo, Grabado - Marzo 1968 | Pablo Picasso | 4609 | 1968 | Lithograph | 30 1/8 x 22 1/4" (76.5 x 56.6 cm) | Peter Stone Collection of Posters by Artists | 464.1976 | Design | Architecture & Design | 1976-10-22 | Y | 7121 | http://www.moma.org/coll
15. Galerie Louise Leiris, Picasso | Pablo Picasso | 4609 | 1971 | Lithograph | 29 13/16 x 19 3/4" (75.8 x 50.2 cm) | Peter Stone Collection of Posters by Artists | 465.1976 | Design | Architecture & Design | 1976-10-22 | Y | 7126 | http://www.moma.org/coll

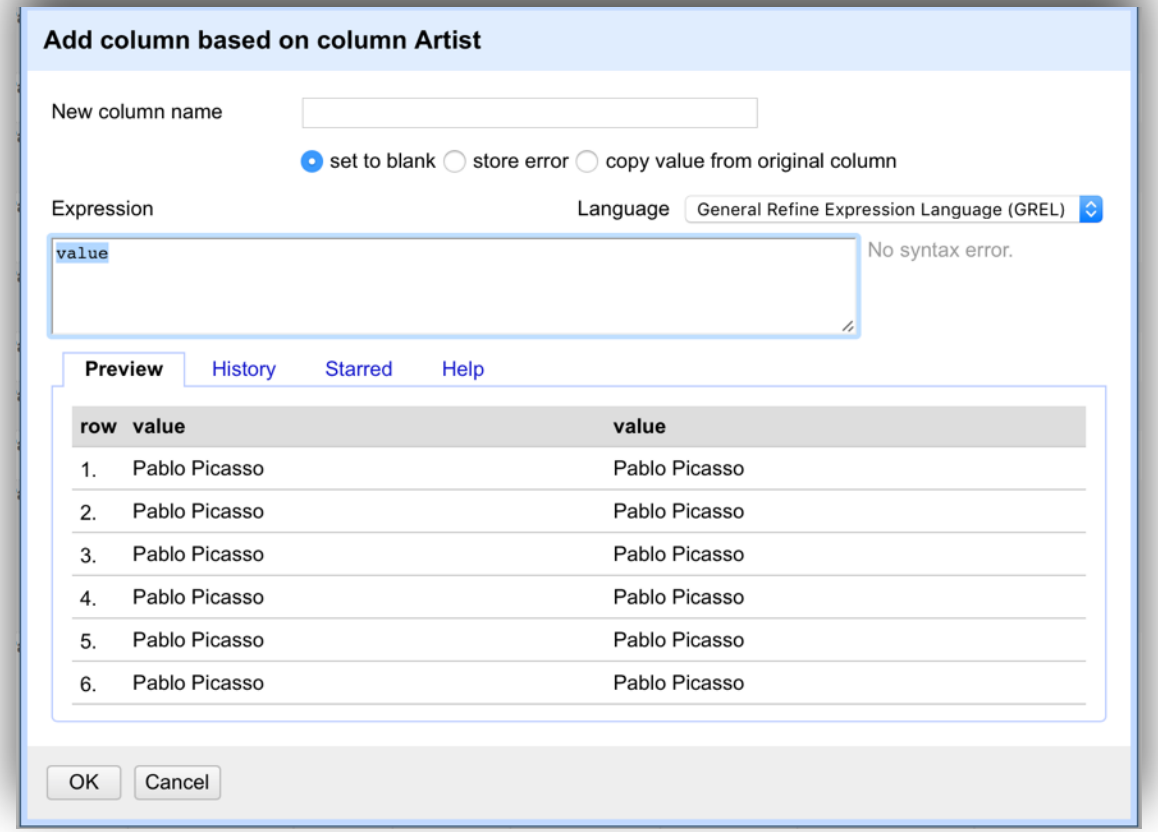


All		Title	Artist	ConstituentID	Date	Medium	Dimensions	CreditLine	AccessionNumb	Classification	Department	DateAcquired	Cataloged	ObjectID	URL
1.	Picasso, "Les Menines," Galerie Louise Leiris	Pablo Picasso	4609	1959	Lithograph	18 3/4 x 26 1/4" (47.6 x 66.6 cm)	Gift of Mourlot Frères	60.1960	Design	Architecture & Design	1960-04-06	Y	4891	http://www.moma.org/coll
2.	Vallauris 1951 Exposition	Pablo Picasso	4609	1951	Linocut	25 5/8 x 19 1/2" (65.1 x 49.5 cm)	Given anonymously	172.1968	Design	Architecture & Design	1968-03-07	Y	5481	http://www.moma.org/coll
3.	Exposition Hispano-Americaine	Pablo Picasso	4609	1951	Lithograph	23 1/4 x 18" (59.1 x 45.7 cm)	Purchase	331.1953	Design	Architecture & Design	1953-04-07	Y	6305	http://www.moma.org/coll
4.	Picasso et le théâtre	Pablo Picasso	4609	1965	Lithograph	35 3/4 x 22 3/4" (90.8 x 57.8 cm)	Gift of James Thrall Soby	331.1966	Design	Architecture & Design	1966-05-10	Y	6308	http://www.moma.org/coll
5.	Exposition Vallauris	Pablo Picasso	4609	1952	Linoleum cut	each: 26 1/8 x 19 3/4" (66.4 x 50.2 cm)	Gift of Curt Valentin	332.1953.1-2	Design	Architecture & Design	1953-04-07	N	6312	
6.	Exposition Vallauris	Pablo Picasso	4609	c.1953	Linocut	each: 37 1/2 x 23 1/4" (95.3 x 59.1 cm)	Gift of Curt Valentin	333.1953.1-2	Design	Architecture & Design	1953-04-07	N	6319	
7.	Exposition Vallauris	Pablo Picasso	4609	c.1953	Poster		Gift of Elizabeth Fuller	334.1953	Design	Architecture & Design	1953-04-07	N	6327	
8.	Toros en Vallauris, 1955	Pablo Picasso	4609	1955	Color woodcut		Gift of Gertrud A. Mellon	335.1957	Design	Architecture & Design	1957-12-04	N	6335	
9.	Vallaurus, Toros	Pablo Picasso	4609	1956	Woodcut	39 x 25 5/8" (99.1 x 65.1 cm)	Gift of Gertrud A. Mellon	336.1957	Design	Architecture & Design	1957-12-04	Y	6342	http://www.moma.org/coll
10.	Vallauris-1952, Exposition	Pablo Picasso	4609	1956	Linocut	39 x 25 3/4" (99.1 x 65.4 cm)	Gift of Gertrud A. Mellon	337.1957	Design	Architecture & Design	1957-12-04	Y	6349	http://www.moma.org/coll
11.	"Congres National/ mouvement de la Paix/ 10 et 11 Mars 1962/ Issy-les-Moulineaux"	Pablo Picasso	4609	1961	Lithograph	39 5/16 x 25 3/16" (99.8 x 64.0 cm)	Peter Stone Collection of Posters by Artists	461.1976	Design	Architecture & Design	1976-10-22	Y	7107	http://www.moma.org/coll
12.	Galerie de l'Elysée, 69 rue du Faubourg St-Honoré, Alex Maguy expose 7 tableaux majeurs de Picasso 30 mai - 30 juin (19)62	Pablo Picasso	4609	1962	Lithograph	25 11/16 x 19" (65.2 x 48.3 cm)	Peter Stone Collection of Posters by Artists	462.1976	Design	Architecture & Design	1976-10-22	Y	7111	http://www.moma.org/coll
13.	Picasso: 60 Years of Graphic Works - Los Angeles County Museum of Art, 25 October - 24 December, 1966	Pablo Picasso	4609	1966	Lithograph	29 x 20 3/16" (73.6 x 51.3 cm)	Peter Stone Collection of Posters by Artists	463.1976	Design	Architecture & Design	1976-10-22	Y	7116	http://www.moma.org/coll
14.	Sala Gaspar - Picasso: Pintura, Dibujo, Grabado - Marzo 1968	Pablo Picasso	4609	1968	Lithograph	30 1/8 x 22 1/4" (76.5 x 56.6 cm)	Peter Stone Collection of Posters by Artists	464.1976	Design	Architecture & Design	1976-10-22	Y	7121	http://www.moma.org/coll
15.	Galerie Louise Leiris, Picasso	Pablo Picasso	4609	1971	Lithograph	29 13/16 x 19 3/4" (75.8 x 50.2 cm)	Peter Stone Collection of Posters by Artists	465.1976	Design	Architecture & Design	1976-10-22	Y	7126	http://www.moma.org/coll

Medium & CreditLine

Faceting

- Key collision based algorithms
  - Fingerprint - most conservative of all of them
  - ngram fingerprint - more flexible, diversity in type of duplicate
  - metaphone 3 - cologne-phonetic. Better check it.
- Nearest neighbour based algorithm
  - They calculate the number of edits between two strings and group them if under a certain threshold. Slower and require more computing power.


Add column based on column Artist
New column name
set to blank
store error
copy value from original column
Expression
Language
General Refine Expression Language (GREL)
value
No syntax error.
Preview
History
Starred
Help
row
value
value
1. Pablo Picasso Pablo Picasso
2. Pablo Picasso Pablo Picasso
3. Pablo Picasso Pablo Picasso
4. Pablo Picasso Pablo Picasso
5. Pablo Picasso Pablo Picasso
6. Pablo Picasso Pablo Picasso
OK
Cancel

DateAcquired

Change date format

```
toString(toDate(value),"dd/MM/yyyy")
```

Date 1/4 - Jython

Date - Add column based on this column—> “object_creation_date_start”

```
import re
pattern = re.compile(r"\d{4}")
return pattern.findall(value)[0]
```

Date - Add column based on this column—> “object_creation_date_end”

```
import re
pattern = re.compile(r"\d{4}|(?<=-)\d{2}")
return pattern.findall(value)[-1]
```

Date 2/4 - GREL

end_date -> add “19” at the beginning of the date

```
return "19" + value
```

“object_creation_date_start” & “object_creation_date_end”

```
value.toDate()
```

end_date -“Add column based on this column—> “diff”

```
diff(cells["end_date"].value,cells["start_date
"].value, "years")
```

Date 3/4 - GREL

Add column based on this column—> “start_date2”

```
value
```

object_creation_date_start -“Transform”

```
cells["end_date"].value
```

object_creation_date_end -“Transform”

```
cells[“start_date2”].value
```

Date 4/4 - GREL

date to string yyyy

```
toString(toDate(value),"yyyy")
```

Date -> remove column

Reconciliation

Compare value in my dataset with value from an external source. If they match link them and extract information

Wikidata

- Wikimedia knowledge base
- Single source of structured information for wiki*
- Organised with unique ID and attribute-value pair *<attribute name, value>*

Reconciliation Wikidata

Artist -“Add column based on this column—> “Wikidata_id”

```
cell.recon.match.id
```

Artist -“Add column based on this column—>

```
"https://www.wikidata.org/wiki/" +
cell.recon.match.id
```

Reconciliation Wikidata

Artist -"Add column based on this column—>

```
cell.recon.match.name
```

Artist -"Add column based on this column—>

```
cell.recon.match.type
```

In this case "Q5" -> a human type in Wikidata

Adding information from Wikidata

Date of birth —> in Wikidata => P569

https://tools.wmflabs.org/openrefine-wikidata/en/api

Adding information from Wikidata

"Artist_id" - Add column by fetching URL ->

```
'https://tools.wmflabs.org/openrefine-wikidata/en/fetch_values?item=' + value + '&prop=P569&label=true'
```

```
value.parseJson().values
```

```
value.replace('[','').replace(']','').replace('"','').replace('+','')
```

```
toString(toDate(value),"dd/MM/yyyy")
```

VIAF

- International Authority File from OCLC
- Link diverse national authority file to a single virtual one
- A VIAF record gives access to all the national records

Reconciliation VIAF

http://localhost:8080/reconcile/viaf

Reconciliation with just one source

http://localhost:8080/reconcile/viaf/JPG (ULAN)

Retrieve IDs of a specific source

http://localhost:8080/reconcile/viafproxy/LC

```
cell.recon.match.id
```

Cell Cross

ColumnName -> **Place**

Add value from another OpenRefine project.

```
cell.cross("Picasso_place_ITN_Workshop", "year")
[0].cells["City"].value
```

GEONAMES

- Geographical database (WGS84)
- >10,000,000 geographical names
- Stable URI and accessible free of charge

Reconciliation Geonames

http://localhost:5000/reconcile

ColumnName -> creation_geonames_id

URI `cell.recon.match.id`

Coordinates

```
replace(substring(cell.recon.match.name,
indexOf(cell.recon.match.name, ' | ')), ' | ', '')
```

Name + Coordinates `cell.recon.match.name`

Add museum reference 1/4

Add column based on this column—> “Museum”

value = **'modern museum of art'**

Reconcile to Wikidata —> Art museum

Museum -“Add column based on this column—> “museum_id”

```
cell.recon.match.id
```

Add museum reference 2/4

Museum -“Add column based on this column—> “Museum_Address”

```
'https://tools.wmflabs.org/openrefine-wikidata/en/fetch_values?item=' + value + '&prop=P969&label=true'
```

Parse JSON

```
value.parseJson().values
```

Add museum reference 3/4

Replace characters

```
value.replace('[','').replace(']','').replace('"'
,'')
```

Add column by Fetching URL -> temp

```
"http://maps.google.com/maps/api/geocode/json?
sensor=false&address=" + escape(value, "url")
```

Add museum reference 4/4

Create a new column -> Museum_lat+long

```
with(value.parseJson().results[0].geometry.locati
on, pair, pair.lat +", " + pair.lng)
```

test - Create a new column -> Latitude

```
value.parseJson().results[0].geometry.location.lat
```

test - Create a new column -> Longitude

```
value.parseJson().results[0].geometry.location.lng
```

OR Add museum reference 1/1

Check the API

```
'http://maps.google.com/maps/api/geocode/json?
sensor=false&address=' + escape(value, 'url')
```

Create a new column

```
with(value.parseJson().results[0].geometry.locati
on, pair, pair.lat +", " + pair.lng)
```

Art and Architecture thesaurus

Medium - Add column by fetching URL—> “AAT”

```
"http://leduc.gamsau.archi.fr/Skosmos/rest/v1/
aat/search?query=" + value + "*&lang=en"
```

Parse JSON -> Transform + Concatenation

```
value.parseJson().results
```

```
forEach(value.parseJson(),v,
[v.uri,v.prefLabel].join('||')).join('::')
```

Art and Architecture thesaurus

Check first value

```
value.split("::")[0]
```

Split column

Exporting in XML?

CSV to XML

https://shancarter.github.io/mr-data-converter/

Other Datasets