

Reconciliation Example Using OpenRefine with the Getty Vocabularies LOD

Developed by Gregg Garcia

J. Paul Getty Trust

ggarcia@getty.edu

Load Excel spreadsheet, JSON file, CSV file, XML file, etc. into OpenRefine and create a new project

The screenshot shows the OpenRefine web interface. At the top, the browser address bar shows '127.0.0.1:3333'. The OpenRefine logo and tagline 'A power tool for working with messy data.' are visible. The main area is titled 'Create Project' and shows a project named 'TMSFields.xlsx'. Below this, a table displays a preview of the data with three columns: 'Column 1', 'Column 2', and 'Column 3'. The table contains 12 rows of data. Below the table, the 'Parse data as' section is active, showing options for 'Excel files', 'JSON files', 'Line-based text files', 'CSV / TSV / separator-based files', 'Fixed-width field text files', 'PC-Axis text files', 'MARC files', 'RDF/N3 files', and 'XML files'. The 'Excel files' option is selected, and the 'Worksheets to Import' section shows 'Sheet1' with 546 rows. There are also checkboxes for 'Ignore first', 'Parse next', 'Discard initial', and 'Load at most' with corresponding line/row counts. On the right side, there are checkboxes for 'Store blank rows', 'Store blank cells as nulls', and 'Store file source (file names, URLs) in each row'. An 'Update Preview' button is located at the bottom right of the parsing options section. The left sidebar contains navigation links for 'Open Project', 'Import Project', and 'Language Settings'. The bottom left corner shows the version 'Version 2.6-rc.2 [TRUNK]' and links for 'Help' and 'About'.

	Column 1	Column 2	Column 3
1.	Sling-bullets		1551880
2.	Finials	Art & Architecture Thesaurus®	1551895
3.	Scaraboids		1551845
4.	Architraves	Art & Architecture Thesaurus®	1551851
5.	Xylophones	Art & Architecture Thesaurus®	1551862
6.	Ephemera	Art & Architecture Thesaurus®	1547300
7.	Akroteria	Beazley Archive Dictionary	1547799
8.	Chalices	GettyGuide glossary term	1551735
9.	Staters	Art & Architecture Thesaurus®	1551749
10.	Medallions	GettyGuide glossary term	1551779
11.	Sestertii	Art & Architecture Thesaurus®	1551796
12.	Herte		1551798

From the drop down of the column you want to reconcile, go to “Edit Column” and select “Add column by fetching URLs”

The screenshot shows the Refine interface with a data table. A context menu is open over the 'Edit column' option for 'Column 2'. A blue arrow points from the text above to the 'Add column by fetching URLs...' option in the menu.

127.0.0.1:3333/project?project=2378633279447

Refine^{OPEN} TMSFields.xlsx Permalink

Facet / Filter Undo / Redo

546 rows

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

All	Column 1	Column 2	Column 3
1. Facet			1551880
2. Text filter	chitecture Thesaurus@		1551895
3. Edit cells			1551845
4. Edit column	chitecture Thesaurus@		1551851
5. Transpose			
6. Sort...			
7. View			
8. Reconcile			

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](#)

- Facet
- Text filter
- Edit cells
- Edit column**
 - Split into several columns...
 - Add column based on this column...
 - Add column by fetching URLs...**
 - Rename this column
 - Remove this column
 - Move column to beginning
 - Move column to end
 - Move column left
 - Move column right
- Transpose
- Sort...
- View
- Reconcile

Refer to sample GREL code with SPARQL query template on Vocab LOD sample queries page (section 3.13)

The screenshot shows the 'Getty Vocabularies: LOD' website. The left sidebar contains a navigation menu with sections 2 and 3. Section 3, 'Getting Information', includes item 3.13 'OpenRefine Reconciliation Service', which is highlighted. The main content area shows a SPARQL query interface with a 'Query:' label and a text box containing the number '1'. Below the text box are two checkboxes: 'Include inferred' (checked) and 'Expand results over equivalent URIs' (unchecked). A blue 'Submit' button is located to the right of the checkboxes.

3.13 OpenRefine Reconciliation Service

[OpenRefine](#) (formerly Google Refine) is a popular and powerful tool for working with messy tabular data: cleaning it; transforming it (including to LOD); extending it with web services; linking it to structured databases. It was originally used for populating Freebase, then open sourced by Google. DERI created some [useful extensions](#): Reconcile & interlink, Export RDF. [LODRefine](#) is a repackaging of these extensions, adding reconciliation against DBpedia, Crowd-sourcing, and Statistics. It was popularized for use by GLAM professionals by Ruben Verborgh, Seth Holland and Max De Wilde through the sites <http://openrefine.org/> and <http://freeyourmetadata.org/>.

A [question has been asked](#) whether GVP LOD can be used as an OpenRefine reconciliation service. The DERI extension includes a "SPARQL full-text search-based Reconciliation" that unfortunately cannot be used, because there's no way to specify that the `luc:term` index should be used (see [issue/33](#)). Nevertheless, one can use the GVP SPARQL service by querying for a fixed label (similar to [Find Subject by Exact English PrefLabel](#)), getting JSON format and parsing the result. Inge van Stokkom of the Rijksmuseum [worked out a detailed solution](#). We reproduce it here with a few changes. Assume you have NL labels and you want to look them up in AAT and fetch the AAT identifier and the EN `prefLabelGVP`:

- Create a column by fetching a URL based on the column that contains the terms

```
'http://vocab.getty.edu/sparql.json?query=select+distinct+*{?x+skos:inScheme+aat.; (x1:prefLabel|x1:altLabel) /gvp:term' + escape(value, 'url') + '"@nl}'
```

- Parse the JSON to obtain the URL:

```
value.parseJson().results.bindings[0].x.value
```

- Parse the identifier out of the URL by adding a column based on this column:

```
value[27, 37]
```

Put GREL code with embedded SPARQL query into expression box, define the new column name and press “OK” button to start process

The screenshot shows a web application interface with a dialog box titled "Add column by fetching URLs based on column Column 1". The dialog box contains the following elements:

- New column name:** A text input field containing "Reconcile1".
- Throttle delay:** A text input field containing "5000" milliseconds.
- On error:** Radio buttons for "set to blank" (selected) and "store error".
- Formulate the URLs to fetch:** A section with a "Language" dropdown set to "General Refine Expression Language (GREL)".
- Expression:** A text area containing the GREL code: `'http://vocab.getty.edu/sparql.json?query=select+distinct*{?x+skos:inScheme+aat;:(x1:prefLabel|x1:altLabel)/gvp:term"' + escape(toLowercase(trim(value)), 'url') + '"@en}'`. A "No syntax error." message is displayed to the right.
- Preview:** A table showing the results of the GREL expression. The table has columns "row", "value", and "url".

row	value	url
1.	Sling-bullets	http://vocab.getty.edu/sparql.json?query=select+distinct*{?x+skos:inScheme+aat;:(x1:prefLabel x1:altLabel)/gvp:term"sling-bullets"@en}
2.	Finials	http://vocab.getty.edu/sparql.json?query=select+distinct*{?x+skos:inScheme+aat;:(x1:prefLabel x1:altLabel)/gvp:term"finials"@en}
3.	Scaraboids	http://vocab.getty.edu/sparql.json?query=select+distinct*{?x+skos:inScheme+aat;:(x1:prefLabel x1:altLabel)/gvp:term"scaraboids"@en}

At the bottom of the dialog box are "OK" and "Cancel" buttons.

OpenRefine displays progress of reconciliation process

The screenshot shows the OpenRefine web interface. At the top, the browser address bar displays the URL `127.0.0.1:3333/project?project=1844835252820`. The OpenRefine logo and the file name `TMSFields.xlsx` are visible. A yellow tooltip box is overlaid on the interface, containing the following text:

```
Create column Reconcile1 at index 1 by fetching URLs based on column Column 1 using expression  
grel:'http://vocab.getty.edu/sparql.json?query=select+distinct*{?x+skos:inScheme+aat;(xl:prefLabel|xl:altLabel)/gvp:term'" + escape(toLowercase(trim(value)), 'url') + "'@en}'
```

Below the tooltip, a table displays 10 rows of data. The table has columns for row number, term, source, and a numerical value. The progress indicator shows `0% complete` with a `Cancel` button. A sidebar on the left contains a section titled `Using facets and filters` with instructions and a link to `Watch these screencasts`.

Row	Term	Source	Value
1.	Sling-bullets		1551880
2.	Finials	Art & Architecture Thesaurus®	1551895
3.	Scaraboids		1551845
4.	Architraves	Art & Architecture Thesaurus®	1551851
5.	Xylophones	Art & Architecture Thesaurus®	1551862
6.	Ephemera	Art & Architecture Thesaurus®	1547300
7.	Akroteria	Beazley Archive Dictionary	1547799
8.	Chalices	GettyGuide glossary term	1551735
9.	Staters	Art & Architecture Thesaurus®	1551749
10.	Medallions	GettyGuide glossary term	1551779

New column contains JSON of reconciliation results after processing

127.0.0.1:3333/project?project=1844835252820

Refine TMSFields.xlsx Permalink Open...

Facet / Filter Undo / Redo 1

10 rows

Show as: rows records Show: 5 10 25 50 rows « first < previous

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](#)

	All	Column 1	Reconcile1	Col
1.	Sling-bullets		{ "head": { "vars": ["x"] }, "results": { "bindings": [] } }	
2.	Finials		{ "head": { "vars": ["x"] }, "results": { "bindings": [{ "x": { "type": "uri", "value": "http://vocab.getty.edu/aat/300002280" } }] } }	Art & Architec Thesaur
3.	Scaraboids		{ "head": { "vars": ["x"] }, "results": { "bindings": [{ "x": { "type": "uri", "value": "http://vocab.getty.edu/aat/300265272" } }] } }	
4.	Architraves		{ "head": { "vars": ["x"] }, "results": { "bindings": [{ "x": { "type": "uri", "value": "http://vocab.getty.edu/aat/300001780" } }, { "x": { "type": "uri", "value": "http://vocab.getty.edu/aat/300298870" } }] } }	Art & Architec Thesaur
5.	Xylophones		{ "head": { "vars": ["x"] }, "results": { "bindings": [{ "x": { "type": "uri", "value": "http://vocab.getty.edu/aat/300041976" } }] } }	Art & Architec Thesaur
6.	Ephemera		{ "head": { "vars": ["x"] }, "results": { "bindings": [{ "x": { "type": "uri", "value": "http://vocab.getty.edu/aat/300028881" } }, { "x": { "type": "uri", "value": "http://vocab.getty.edu/aat/300264821" } }] } }	Art & Architec Thesaur
7.	Akroteria		{ "head": { "vars": ["x"] }, "results": { "bindings": [] } }	Beazley Archive Dictionar
8.	Chalices		{ "head": { "vars": ["x"] }, "results": { "bindings": [{ "x": { "type": "uri", "value": "http://vocab.getty.edu/aat/300194762" } }] } }	GettyGu glossary
9.	Staters		{ "head": { "vars": ["x"] }, "results": { "bindings": [{ "x": { "type": "uri", "value": "http://vocab.getty.edu/aat/300191666" } }] } }	Art & Architec Thesaur
10.	Medallions		{ "head": { "vars": ["x"] }, "results": { "bindings": [{ "x": { "type": "uri", "value": "http://vocab.getty.edu/aat/300077354" } }, { "x": { "type": "uri", "value": "http://vocab.getty.edu/aat/300077357" } }] } }	GettyGu glossary

In the drop down for the new column, select “Transform”

127.0.0.1:3333/project?project=1844835252820

Refine TMSFields.xlsx Permalink

Facet / Filter Undo / Redo 1

10 rows

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

All	Column 1	Reconcile1
1.	Sling-bullets	Facet
2.	Finials	Text filter
3.	Scaraboids	Edit cells
4.	Architraves	Edit column
5.	Xylophones	Transpose
6.	Ephemera	Sort...
7.	Akroteria	View
8.	Chalices	Reconcile
9.	Staters	
10.	Medallions	

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](#)

Transform...

Common transforms

Fill down

Blank down

Split multi-valued cells...

Join multi-valued cells...

Cluster and edit...

Put GREL code from Vocabularies sample queries page into the expression box for new column transformation

The screenshot shows a web application interface with a dialog box titled "Custom text transform on column Reconcile1". The dialog has a title bar and a main content area. At the top, it says "Expression" and "Language" (General Refine Expression Language (GREL)). Below that is a text input field containing the GREL expression: `value.parseJson().results.bindings[0].x.value`. To the right of the input field, it says "No syntax error." Below the input field are tabs for "Preview", "History", "Starred", and "Help". The "Preview" tab is active, showing a table with three rows. The first row has an error: "Error: Cannot retrieve field from null". The second and third rows show the transformed values: `http://vocab.getty.edu/aat/300002280` and `http://vocab.getty.edu/aat/300265272`. At the bottom of the dialog, there are options for "On error": "keep original" (selected), "set to blank", and "store error". There is also a checkbox for "Re-transform up to 10 times until no change". At the very bottom are "OK" and "Cancel" buttons.

127.0.0.1:3333/project?project=1844835252820

Refine TMSFields.xlsx Permalink

Facet / Filter Undo / Redo 1

10 rows

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

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

Custom text transform on column Reconcile1

Expression Language General Refine Expression Language (GREL)

`value.parseJson().results.bindings[0].x.value` No syntax error.

Preview History Starred Help

row	value	value.parseJson().results.bindings[0].x.value
1.	<code>{ "head" : { "vars" : ["x"] }, "results" : { "bindings" : [] } }</code>	Error: Cannot retrieve field from null
2.	<code>{ "head" : { "vars" : ["x"] }, "results" : { "bindings" : [{ "x" : { "type" : "uri", "value" : "http://vocab.getty.edu/aat/300002280" } }] } }</code>	<code>http://vocab.getty.edu/aat/300002280</code>
3.	<code>{ "head" : { "vars" : ["x"] }, "results" : { "bindings" : [{ "x" : { "type" : "uri", "value" : "http://vocab.getty.edu/aat/300265272" } }] } }</code>	<code>http://vocab.getty.edu/aat/300265272</code>

On error keep original Re-transform up to 10 times until no change
 set to blank
 store error

OK Cancel

Newly transformed column will display URIs for reconciled values; Notice examples “Akroteria” and “Sling-bullets” did not reconcile

127.0.0.1:3333/project?project=1844835252820

Refine TMSFields.xlsx Permalink

Text transform on 8 cells in column Reconcile1:
grel:value.parseJson().results.bindings[0].x.value Undo Open...

Facet / Filter Undo / Redo 2

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](#)

10 rows


Show as: rows records Show: 5 10 25 50 rows « first < previous

All	Column 1	Reconcile1	Column 2	Column 3
1.	Sling-bullets	{ "head" : { "vars" : ["X"], "results" : { "bindings" : [] } }		1551880
2.	Finials	http://vocab.getty.edu/aat/300002280	Art & Architecture Thesaurus®	1551895
3.	Scaraboids	http://vocab.getty.edu/aat/300265272		1551845
4.	Architraves	http://vocab.getty.edu/aat/300001780	Art & Architecture Thesaurus®	1551851
5.	Xylophones	http://vocab.getty.edu/aat/300041976	Art & Architecture Thesaurus®	1551862
6.	Ephemera	http://vocab.getty.edu/aat/300028881	Art & Architecture Thesaurus®	1547300
7.	Akroteria	{ "head" : { "vars" : ["X"], "results" : { "bindings" : [] } }	Beazley Archive Dictionary	1547799
8.	Chalices	http://vocab.getty.edu/aat/300194762	GettyGuide glossary term	1551735
9.	Staters	http://vocab.getty.edu/aat/300191666	Art & Architecture Thesaurus®	1551749
10.	Medallions	http://vocab.getty.edu/aat/300077354	GettyGuide glossary term	1551779

Investigate why some values did not reconcile;
“Akroteria” did not reconcile because search term was made all lowercase and AAT record has term with first letter capitalized; Maybe try again using case insensitive search

 Research

[Research Home](#) ▶ [Tools](#) ▶ [Art & Architecture Thesaurus](#) ▶ [Search Results](#)

 **Art & Architecture Thesaurus® Online**
Search Results


[New Search](#) [Previous Page](#) [Help](#)


Find Name: **Akroteria**

Logic:

Note: 1 result

[View Selected Records](#) [Select All Records](#) [Clear All](#) [First](#) [Previous](#) [Next](#) [Last](#)
Page: **1**

Click the  icon to view the hierarchy.
Check boxes to view multiple records at once.

1.  **acroteria**
(<culminating and edge ornaments for architectural>, architectural elements, ... Components
(hierarchy name)) [300002214]
Akroteria

[New Search](#) [First](#) [Previous](#) [Next](#) [Last](#)
Page: **1**

In some cases, like “Sling bullets”, the term may not exist in the Getty Vocabularies

The screenshot shows a web browser window with the address bar containing the URL: www.getty.edu/vow/AATServlet?english=N&find=Sling+bullets&logic=AND&page=1¬e=. The page header includes the Getty Research logo and the text "Research". Below the header, there is a breadcrumb trail: "Research Home ▶ Tools ▶ Art & Architecture Thesaurus ▶ Search Results". The main title is "Art & Architecture Thesaurus® Online Search Results". Navigation links include "New Search" (with a magnifying glass icon), "Previous Page" (with a left arrow icon), and "Help" (with a question mark icon). The search input field shows "Find Name: Sling bullets". Below this, the fields "Logic:" and "Note:" are visible, with "0 results" displayed on the right side. A message states: "Your search has produced *no* results. Please refine your search by clicking on [New Search](#)." At the bottom, there is another "New Search" button.