

OpenRefine: Pulling data from Ideas Exchange

Main items for today

- 1) Basic setup of creating a project in OpenRefine and using Fetch URLs
- 2) Parsing through Html
- 3) Extracting and applying JSON to make script reusable
- 4) Brief troubleshooting example

1. Gather URLs

Page through IE pages and establish pattern for page URLs. Ex:

<https://ideas.exlibrisgroup.com/forums/935112-rapido?filter=top&page=1>

<https://ideas.exlibrisgroup.com/forums/935112-rapido?filter=top&page=2>

<https://ideas.exlibrisgroup.com/forums/935112-rapido?filter=top&page=3>

<https://ideas.exlibrisgroup.com/forums/935112-rapido?filter=top&page=4>

2. Create Project > Get Data From > Clipboard

Copy URLs from above and paste into project import window.

A power tool for working with messy data.

New Version Download OpenRefine v3.7.7 now.

Create a project by importing data. What kinds of data files can I import?
TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported

Get data from

This Computer

Web Addresses (URLs)

Clipboard

Databases

Paste data from clipboard here:

<https://ideas.exlibrisgroup.com/forums/935112-rapido?filter=top&page=1>
<https://ideas.exlibrisgroup.com/forums/935112-rapido?filter=top&page=2>
<https://ideas.exlibrisgroup.com/forums/935112-rapido?filter=top&page=3>

Next

3. Configure parsing options

Since we have no headers, use Parse Next 0 lines as column headers.

Remove the predictive “:” from column separation. (Either comma-separated or tab-separated should work for this example.)

Column 1

1. <https://ideas.exlibrisgroup.com/forums/935112-rapido?filter=top&page=1>
2. <https://ideas.exlibrisgroup.com/forums/935112-rapido?filter=top&page=2>
3. <https://ideas.exlibrisgroup.com/forums/935112-rapido?filter=top&page=3>

Parse data as

Character encoding

☐ Disable

CSV / TSV / separator-based files

Line-based text files

Fixed-width field text files

PC-Axis text files

Columns are separated by

☒ commas (CSV)

☐ tabs (TSV)

☐ custom :

☐ Ignore first 0 line(s) at beginning of file

☒ Parse next 0 line(s) as column headers

Name project. Create project.

4. Use URLs to fetch html into a new column

Column 1 > Edit column > Add column by fetching URLs

Now you have all of this hideous html that you have to parse through.

3 rows

Show as: **rows** records Show: 5 10 25 50 100 500 1000 rows

All	Column 1	Fetch HTML
1.	https://ideas.exlibrisgroup.com/forums/935112-rapido?filter=top&page=1	<pre><!DOCTYPE html> <html lang="en" class="ltr"> <head> <meta http-equiv="X-UA-Compatible" content="IE=Edge,chrome=1"> <meta charset="utf-8"> <title>Rapido: Top (63 ideas) – Have an idea for Ex Libris?</title> <meta name="viewport" content="width=device-width, initial-scale=1"> <meta name="csrf-param" content="authenticity_token" /> <meta name="csrf-token" content="Q4DHWqQrvDylWKWa7zPj-FacS6jC0wH4MK1fcb4/ <link rel="canonical" href="https://ideas.exlibrisgroup.com/forums/935112-rapido" /> <script type="text/javascript"> (function(window) { var methods = ['log', 'dir', 'trace'];</pre>

5. parseHtml()

Look through all of the html to find divs, classes, etc. that you want to pull and create new columns for those.

Title

```
forEach(  
  value.parseHtml().select("ol.uvList")[0].select("h2.uvIdeaTitle"),  
  e,  
  e.htmlText()  
).join(" | ")
```

Description

```
forEach(  
  value.parseHtml().select("ol.uvList")[0].select("div.uvIdeaDescription"),  
  e,  
  e.htmlText()  
).join(" | ")
```

VoteCount

```
forEach(  
  value.parseHtml().select("ol.uvList")[0].select("div.uvIdeaVoteCount"),  
  e,  
  e.htmlText()  
).join(" | ")
```

6. Clean up

Remove columns no longer needed

Remove column Fetch Html

Move columns

7. Split rows

[illegible]

Populate all cells in Column 1 (Edit cells > Fill down)

9. Extract the JSON to reuse these steps with other, appropriate data – the BEST thing about this entire process!

Save file as rapido_template.json

<https://ideas.exlibrisgroup.com/forums/929326-esploro?filter=top&page=1>

<https://ideas.exlibrisgroup.com/forums/929326-esploro?filter=top&page=2>

<https://ideas.exlibrisgroup.com/forums/929326-esploro?filter=top&page=3>

<https://ideas.exlibrisgroup.com/forums/929326-esploro?filter=top&page=4>

<https://ideas.exlibrisgroup.com/forums/929326-esploro?filter=top&page=5>

<https://ideas.exlibrisgroup.com/forums/929326-esploro?filter=top&page=6>

Everything runs, but oh no, the number of rows is off!

Looking at the data, things go off the rails at line 46 - Description "Consolidated Task List for all assets requiring staff intervention" is "2nd level unit." Looks suspicious.

Looking at the web page, there a bunch of pipe characters in the description for line 45!

Ok, go back to json file and edit pipe to double-pipe or something else amazingly unique

13. Edit contents of rapido_template.json file and rename/save (e.g., esploro_template.json)

Find/Replace `join("\|\"")` to `join("\||\"")`

Find/Replace `"separator": "|"` with `"separator": "||"`

Copy/paste into Esploro project and apply changes (Perform Operations)

Et voila!