Working with Text II: Cleaning and Curating Messy Data with OpenRefine

Gary Stringer

Digital Humanities Lab, University of Exeter

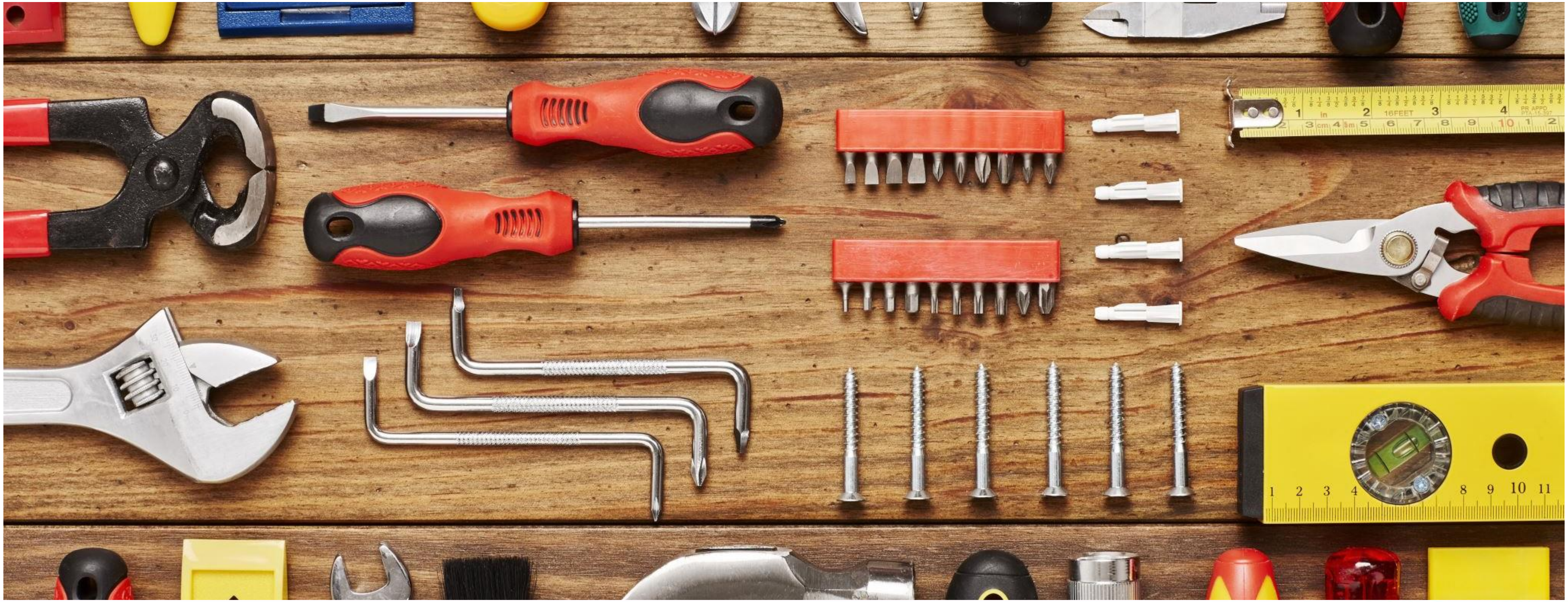
Recap: Messy Data

- We're swamped with data in all kinds of formats
- Manually collected data is often inconsistent and messy
- We often need to:
  - Standardise terms and vocabulary
  - Highlight or extract structure
  - Extract terms or names from general descriptive text
  - Manipulate data into usable structured formats

Recap: Using Regex

- Part one covered a simple tool to find and replace terms and structures, that's commonly built into many applications
- It looked at ways of manipulating your data by finding and replacing specific patterns
- Today, we'll look at a tool that works in similar ways, but adds a few interesting tricks and shortcuts
- It also has the potential to reuse data from outside sources

Building your Toolkit



What is OpenRefine?

- A tool initially developed by Google, but now Open Sourced
- A tool for working with 'messy data'
- An easy way of splitting data (atomisation)
- An easy way of matching, merging and clustering data
- A toolkit for managing consistency and standardisation
- A way of bringing data from other sources (authorities) into your own datasets

Downloading and Installing OpenRefine

Procedure

- Download from: https://openrefine.org/download.html
- Ensure you get at least v 3.5.2
- On Windows, use the Windows Kit if you have installed Java from your institution or IT department
- Otherwise use embedded Java version
- *It's important to keep OpenRefine and Java updated regularly*

Considerations

- Note that OpenRefine runs as a webserver on your computer, but no internet connection is needed
- No data leaves your computer, so it's good for confidentiality
- If you're handling particularly sensitive data, make sure your local firewall blocks the relevant port (3333)

Installing OpenRefine

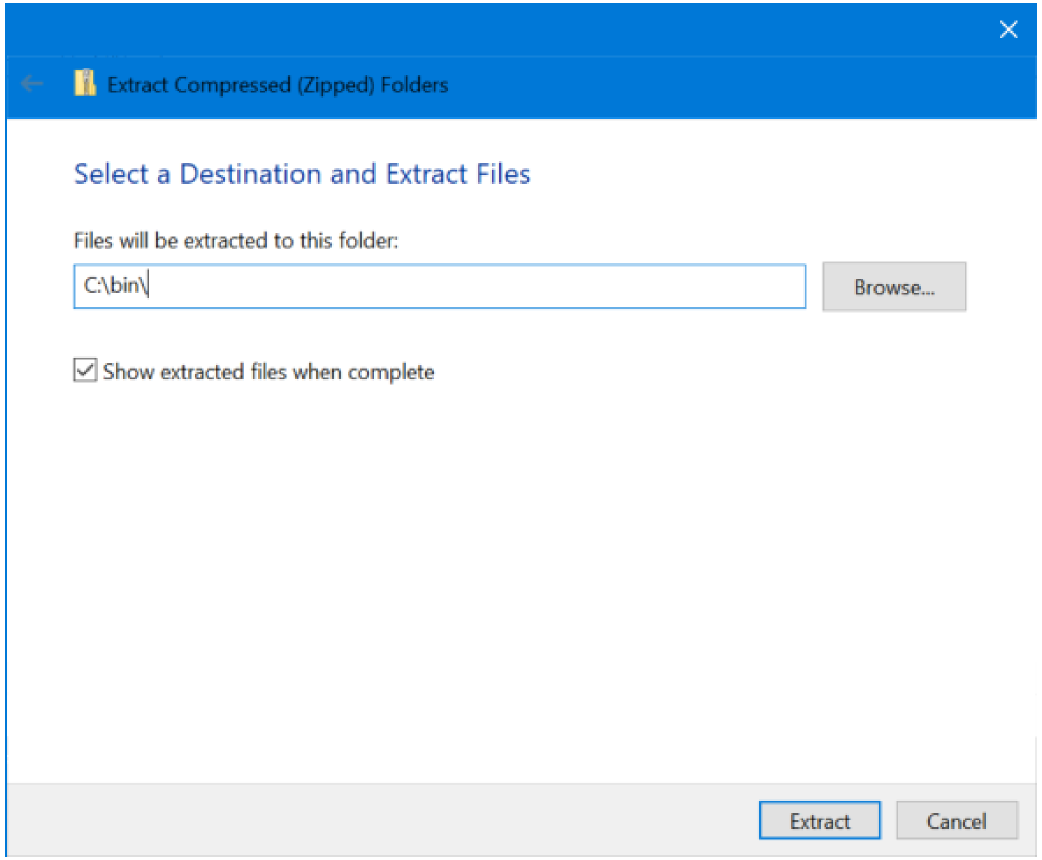
Procedure

- Download the relevant kit for your operating system - this will be a .zip file
- Unpack it to a suitable location
  - If on your hard drive, will be faster, but work needs to be backed up
  - If on cloud storage (e.g. OneDrive), may be slower but is more secure
- Run as per the download instructions (doubleclick openrefine.exe)

Considerations

- The download is 170MB, so ensure you have space to download and install it
- You might want to create a shortcut or link to the startup script
- I tend to unpack OpenRefine into c:\bin\ (windows)
- When you upgrade, the new version will have access to existing projects

Installing OpenRefine (2)



Extract Compressed (Zipped) Folders

Select a Destination and Extract Files

Files will be extracted to this folder:

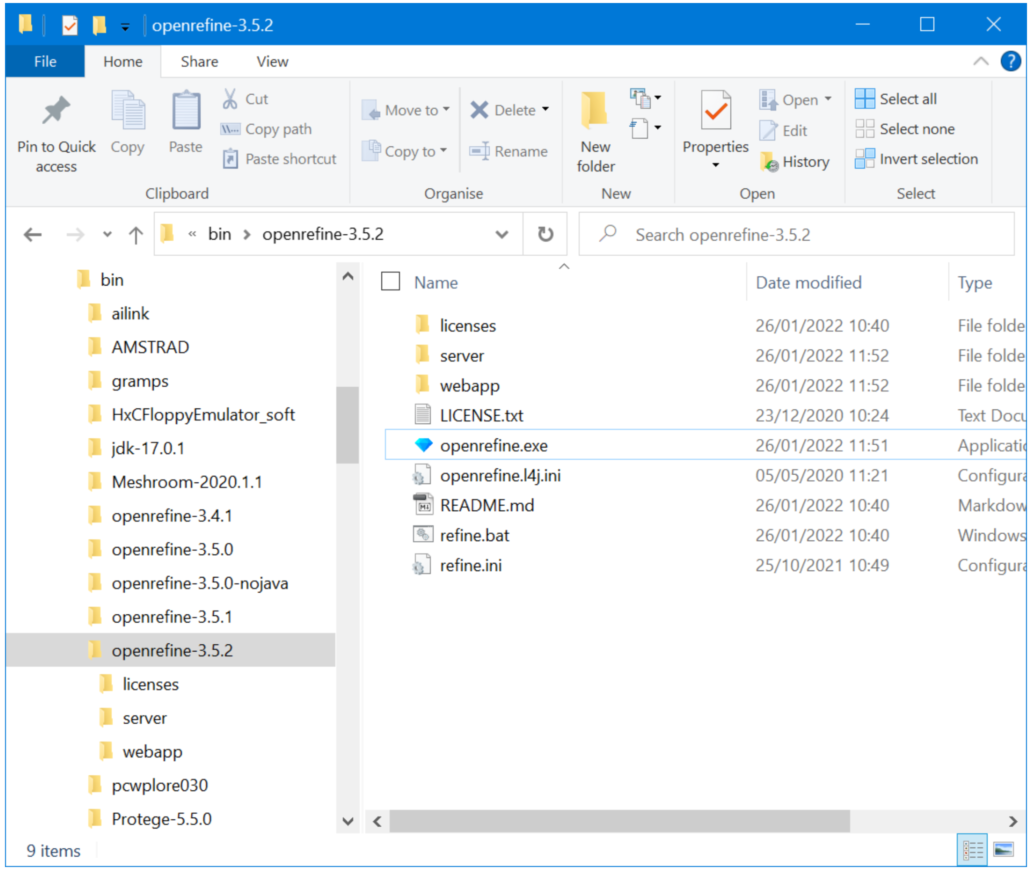
C:\bin\

☑ Show extracted files when complete

Procedure

- In Windows, right click on the downloaded file and choose 'extract all'
- Choose the location and extract
- The archive will unpack with a version number
- Installing new versions will automatically preserve projects

Running the Program



« bin › openrefine-3.5.2

- bin
  - ailink
  - AMSTRAD
  - gramps
  - HxCFloppyEmulator_soft
  - jdk-17.0.1
  - Meshroom-2020.1.1
  - openrefine-3.4.1
  - openrefine-3.5.0
  - openrefine-3.5.0-nojava
  - openrefine-3.5.1
  - openrefine-3.5.2
    - licenses
    - server
    - webapp
  - pcwplore030
  - Protege-5.5.0

Name	Date modified	Type
licenses	26/01/2022 10:40	File folde
server	26/01/2022 11:52	File folde
webapp	26/01/2022 11:52	File folde
LICENSE.txt	23/12/2020 10:24	Text Docu
openrefine.exe	26/01/2022 11:51	Applicati
openrefine.l4j.ini	05/05/2020 11:21	Configura
README.md	26/01/2022 10:40	Markdow
refine.bat	26/01/2022 10:40	Windows
refine.ini	25/10/2021 10:49	Configura

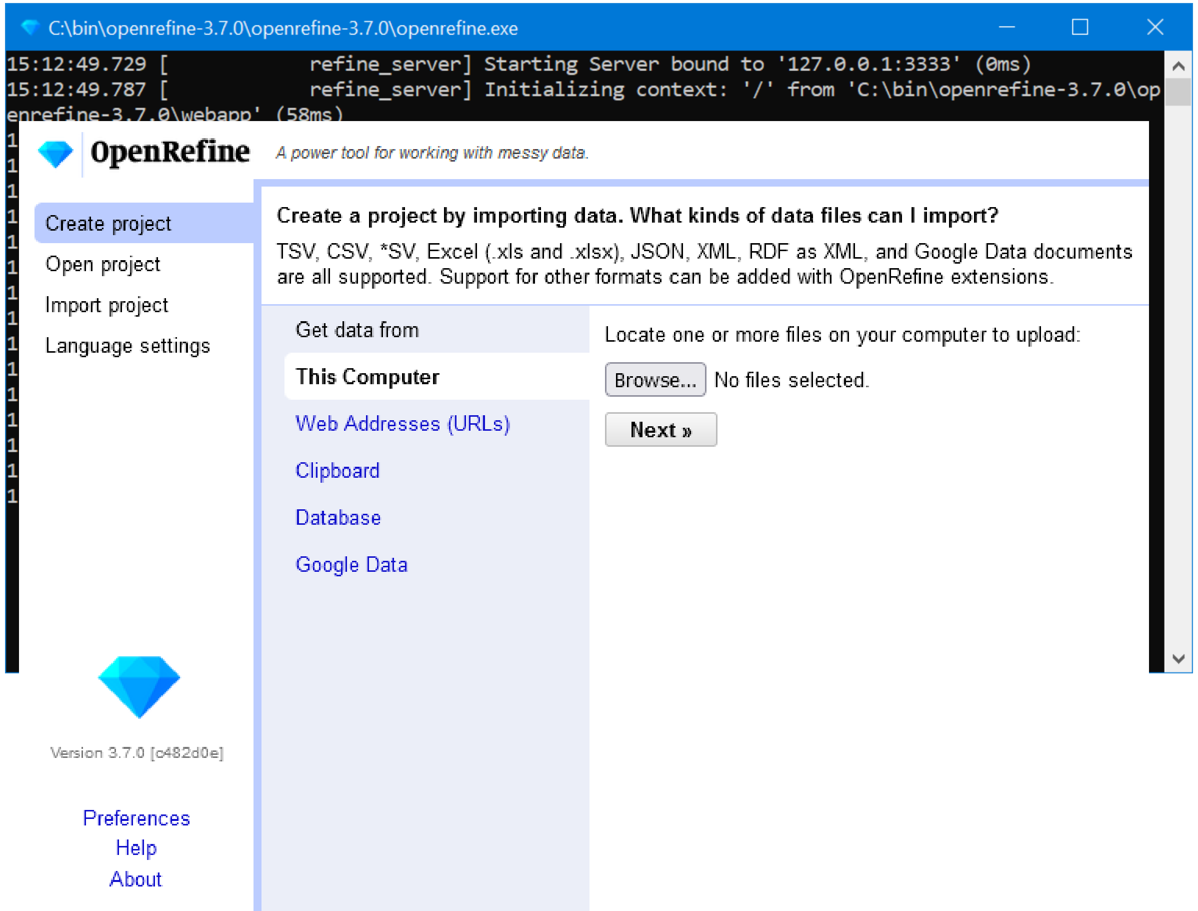
9 items

Finding the executable

- Try the OpenRefine application first
  - If that doesn't work, try the 'refine.bat'
- In Windows, you'll see a black command line box open up
- The OpenRefine interface will then open in a browser window
  - If you don't see it at first, try hunting around in your browser tabs!

Working with OpenRefine

- So, if all is successful, you'll see a black text window
  - any errors will appear there
- And your default browser should open at the OpenRefine interface
  - Note that OpenRefine doesn't work *quite* as well with MS Edge than Chrome or Firefox – you can copy and paste the URL into either of these.



```
15:12:49.729 [          refine_server] Starting Server bound to '127.0.0.1:3333' (0ms)
15:12:49.787 [          refine_server] Initializing context: '/' from 'C:\bin\openrefine-3.7.0\op
enrefine-3.7.0\webapp' (58ms)
```

A power tool for working with messy data.

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. Support for other formats can be added with OpenRefine extensions.

Locate one or more files on your computer to upload:

No files selected.

Version 3.7.0 [c482d0e]

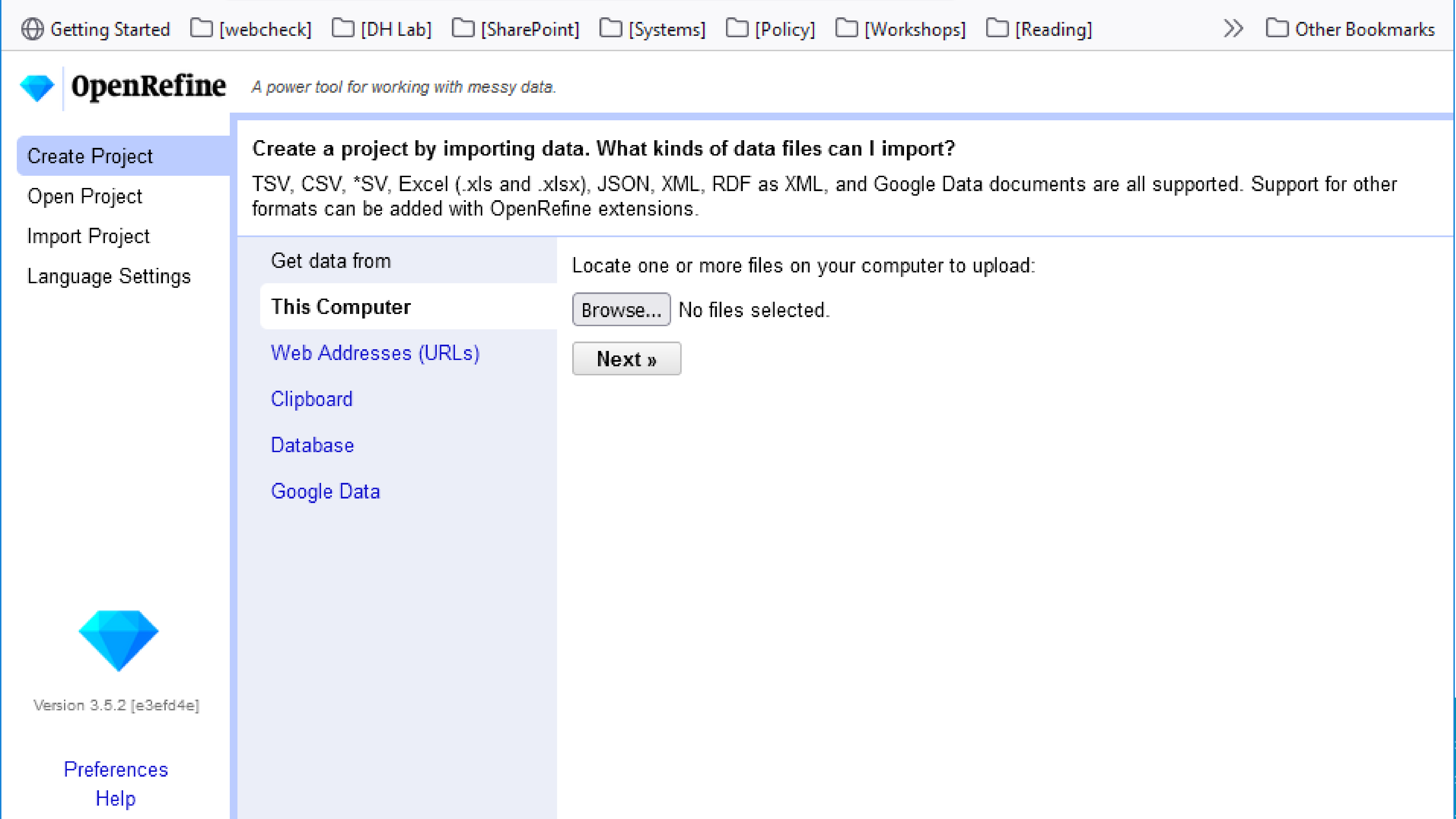
Working with a Dataset

- To represent a common and typical problem that OpenRefine can assist with, we are making some catalogue data available
- This is a direct Excel dump from the CALM catalogue system, but you do not need Excel or CALM to work with it

- You can download sample data from here:
- https://github.com/ExeterDigitalHumanities/openrefine/

Importing your Data

- TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported
- You can also download data directly from the web, or paste data from the clipboard
- If you work with Google Drive, Data or Sheets, there are convenient ways to access these
  - As a Google initiated project, Open Refine was built around these



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. Support for other formats can be added with OpenRefine extensions.

Locate one or more files on your computer to upload:

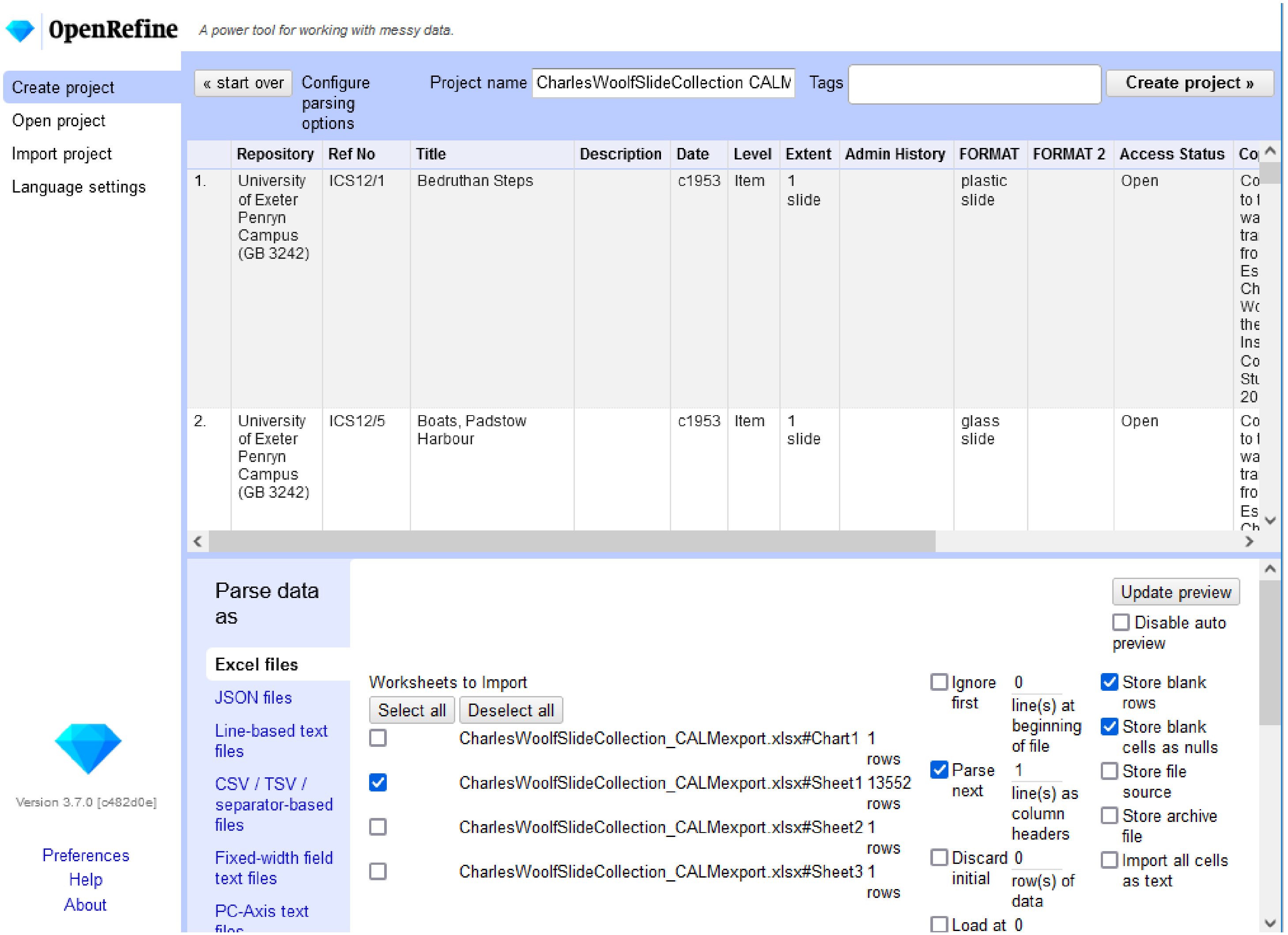
No files selected.

Preview pane

This step shows you a sample of your data, and lets you confirm a few guesses it's made about it

- Give it a clear unambiguous project name
- Typing tags can help to classify projects with identifiers
- Check that you've chosen the correct worksheets or datasets in the lower pane
- Check that any header rows are detected

- When done, click 'Create project >>'



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

Create project
Open project
Import project
Language settings

« start over | Configure parsing options | Project name CharlesWoolfSlideCollection CALM | Tags | Create project »

	Repository	Ref No	Title	Description	Date	Level	Extent	Admin History	FORMAT	FORMAT 2	Access Status	Co
1.	University of Exeter Penryn Campus (GB 3242)	ICS12/1	Bedruthan Steps		c1953	Item	1 slide		plastic slide		Open	Co to t wa tra fro Es Ch Wo the Ins Co Stu 20
2.	University of Exeter Penryn Campus (GB 3242)	ICS12/5	Boats, Padstow Harbour		c1953	Item	1 slide		glass slide		Open	Co to t wa tra fro Es Ch

Parse data as
Excel files
JSON files
Line-based text files
CSV / TSV / separator-based files
Fixed-width field text files
PC-Axis text files

Worksheets to Import
Select all | Deselect all
- □ CharlesWoolfSlideCollection_CALMexport.xlsx#Chart1 1 rows
- ☑ CharlesWoolfSlideCollection_CALMexport.xlsx#Sheet1 13552 rows
- □ CharlesWoolfSlideCollection_CALMexport.xlsx#Sheet2 1 rows
- □ CharlesWoolfSlideCollection_CALMexport.xlsx#Sheet3 1 rows

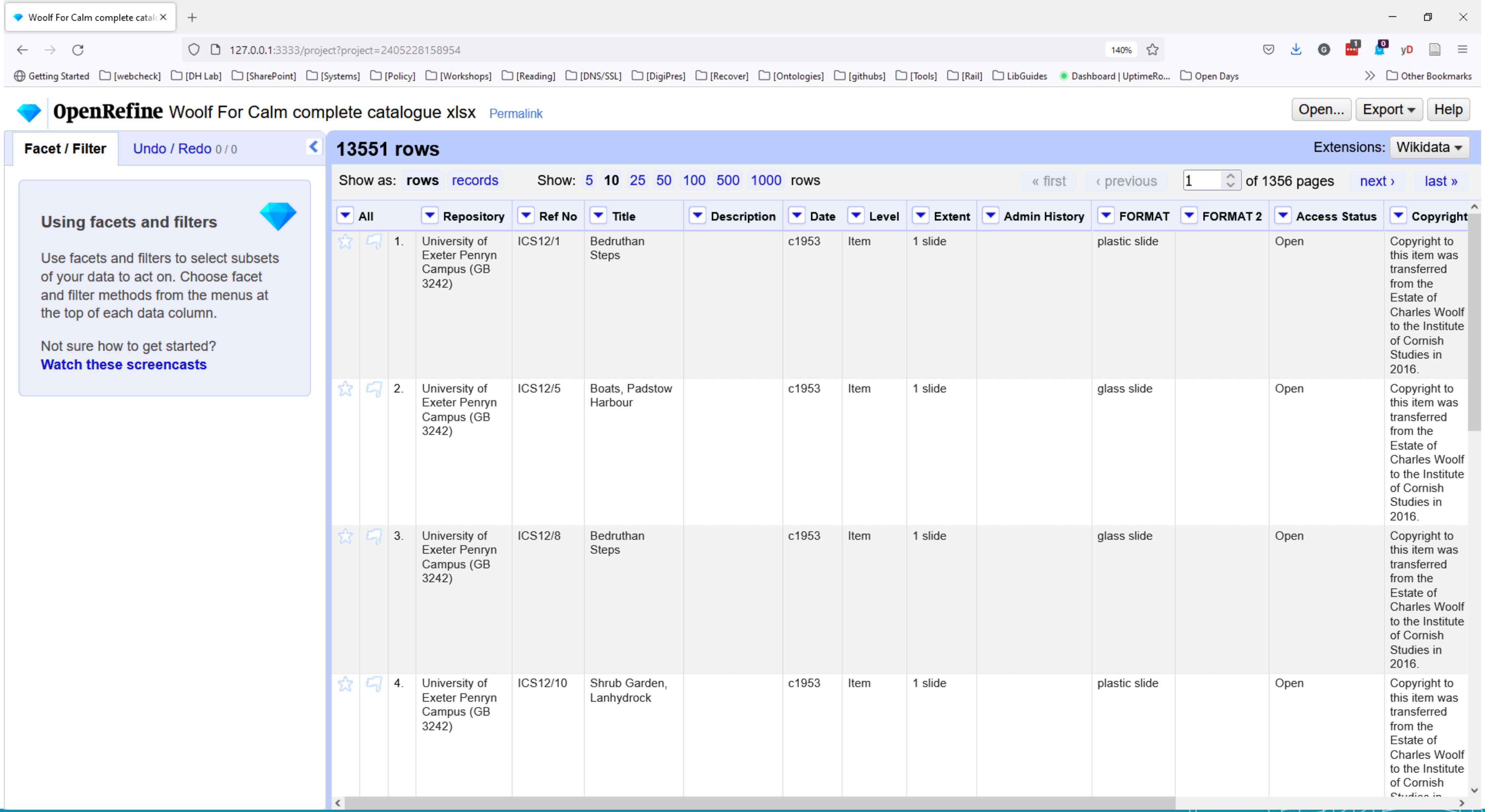
□ Ignore first 0 line(s) at beginning of file
☑ Parse next 1 line(s) as column headers
□ Discard initial 0 row(s) of data
□ Load at 0

☑ Store blank rows
☑ Store blank cells as nulls
□ Store file source
□ Store archive file
□ Import all cells as text

Update preview
□ Disable auto preview

Version 3.7.0 [c482d0e]

Preferences
Help
About



OpenRefine Woolf For Calm complete catalogue xlsx Permalink

13551 rows

Show as: **rows** records   Show: 5 **10** 25 50 100 500 1000 rows   1 of 1356 pages

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	Repository	Ref No	Title	Description	Date	Level	Extent	Admin History	FORMAT	FORMAT 2	Access Status	Copyright
1.	University of Exeter Penryn Campus (GB 3242)	ICS12/1	Bedruthan Steps		c1953	Item	1 slide		plastic slide		Open	Copyright to this item was transferred from the Estate of Charles Woolf to the Institute of Cornish Studies in 2016.
2.	University of Exeter Penryn Campus (GB 3242)	ICS12/5	Boats, Padstow Harbour		c1953	Item	1 slide		glass slide		Open	Copyright to this item was transferred from the Estate of Charles Woolf to the Institute of Cornish Studies in 2016.
3.	University of Exeter Penryn Campus (GB 3242)	ICS12/8	Bedruthan Steps		c1953	Item	1 slide		glass slide		Open	Copyright to this item was transferred from the Estate of Charles Woolf to the Institute of Cornish Studies in 2016.
4.	University of Exeter Penryn Campus (GB 3242)	ICS12/10	Shrub Garden, Lanhydrock		c1953	Item	1 slide		plastic slide		Open	Copyright to this item was transferred from the Estate of Charles Woolf to the Institute of Cornish

Layout of OpenRefine, Rows vs Records

Questions

- How is data organised in OpenRefine?
- How do I access options to amend data in OpenRefine?
- What is the difference between Rows and Records in OpenRefine?
- How do I work with single cells that contain multiple values in a list?

Objectives

- Locate controls for navigating data in OpenRefine
- Find options to work with data through the OpenRefine dropdown menus
- Split cells which contain multiple bits of data so that each piece of data is in its own cell

Faceting and filtering

Definitions

- A **facet** is a field or column in your dataset, or rather, the range of values in a column
- A **filter** allows you to temporarily remove records matching a value in a field
- Facets and fields can help you explore your data, or operate on a subset

Examples

- Choose 'Facet > Text Facet' on the Date column
- You'll see the range of values used to represent a date
- Click 'cluster' to merge and standardise similar values
- Choose 'Text Filter' and filter on '1958'
- You'll see only data for that year

Clustering

Definitions

- Clustering in OpenRefine means to group your data values together by value
- It's achieved by selecting a facet on the drop-down menu for any field
- You can then edit similar values to improve consistency

Examples

- Use clustering to identify and edit or replace varying forms of the same data value with a single consistent value
- You can also use 'Edit Cells > Cluster and Edit' for a more interactive method of doing this in bulk

Working with columns and sorting

Definitions

- Columns can easily moved, reordered or sorted
- Sorting data will sort all records in the chosen order
- Splitting a column creates new columns

Examples

- Use the drop-down on a column header and select 'Edit column'
  - Choose 'Rename' to change the header
  - Choose 'Remove' to delete the column
  - Use 'Move column...' to move it within the sheet
  - Choose 'Join...' to add values in multiple columns together

How might you use these operations to change a separator within a column, e.g. in 'Ref No'?

Introduction to Transformations & GREL

Definitions

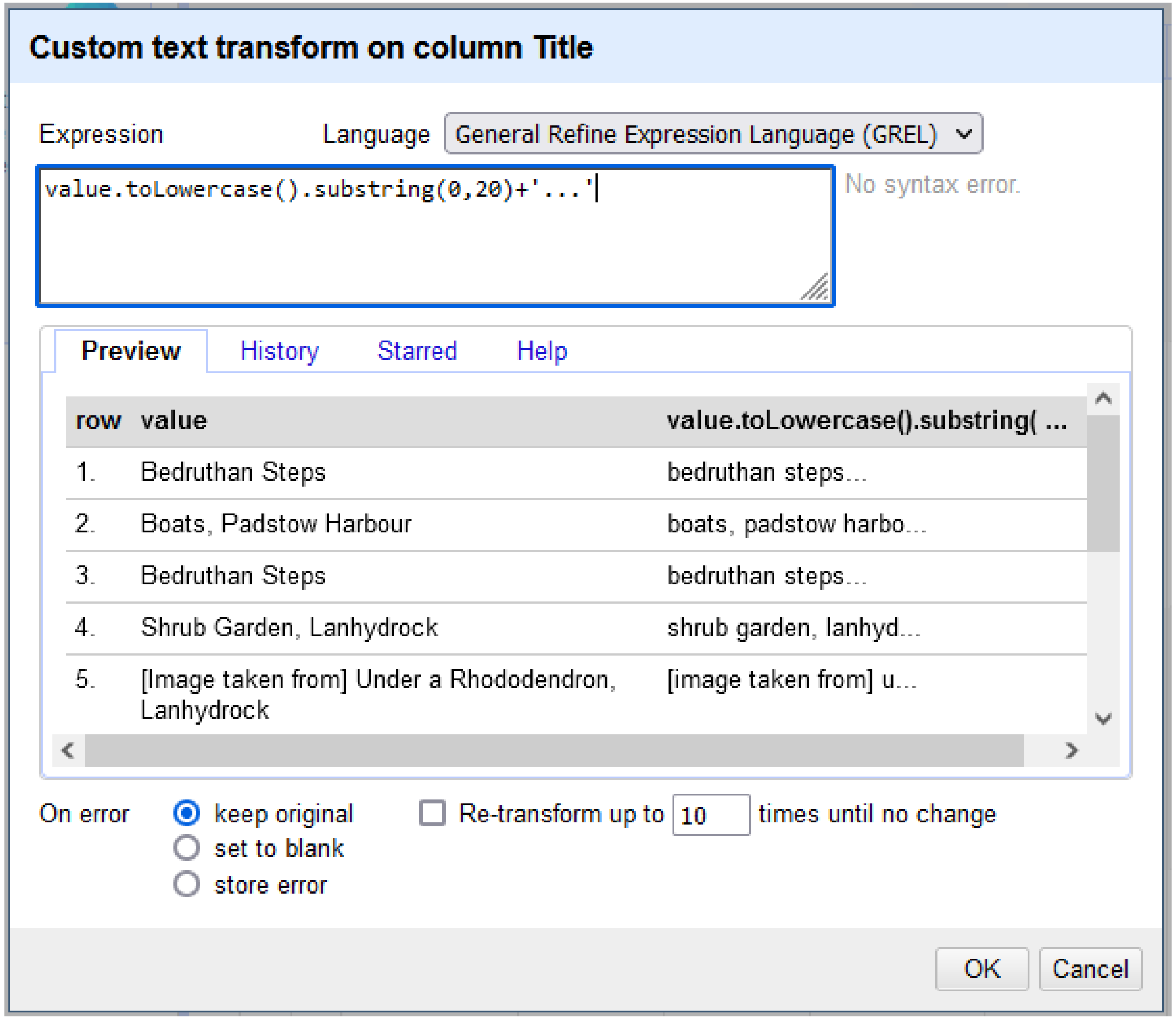
- GREL, the General Refine Expression Language, is a way to do more complex operations on your data
- You can use GREL in transformations of your data
- You can also use regex expressions too
- See the docs for more info:
  - https://openrefine.org/docs/manual/grel

Examples

- On the Title column, select 'Edit cells' then 'Transform'.
- Paste 'value.toLowercase()' in the Expression box
- You should see the results in Preview
- Now try some of the 'Common Transforms' under 'Edit cells'

Writing Transformations

- Transformations can be constructed to do multiple operations
  - Use the dot notation to perform sequential tasks
  - Use the '+' to concatenate string values
  - Note any red 'syntax errors'
- For example, here we are replacing the Title column with a shortened version of the Title, by:
  - Putting it into lower case
  - Shortening it to 20 characters
  - Adding an ellipsis ('...') to show it's abbreviated
- If you need to keep the original Title column, don't forget to first make a copy ('Add column based on this column') of that column.
- The full expression is:
  - value.toLowercase().substring(0,20)+'...’
- Note the case of toLowercase() – it's all case sensitive
- Use the GREL reference for other functions!



Custom text transform on column Title

Expression — Language: General Refine Expression Language (GREL)

```
value.toLowercase().substring(0,20)+'...'
```

No syntax error.

row	value	value.toLowercase().substring(...
1.	Bedruthan Steps	bedruthan steps...
2.	Boats, Padstow Harbour	boats, padstow harbo...
3.	Bedruthan Steps	bedruthan steps...
4.	Shrub Garden, Lanhydrock	shrub garden, lanhyd...
5.	[Image taken from] Under a Rhododendron, Lanhydrock	[image taken from] u...

On error: ● keep original ○ set to blank ○ store error

□ Re-transform up to 10 times until no change

Transformations - Undo and Redo

Definitions

- For each operation you make, OpenRefine records the action
- All operations are reversible, so you can step back and forth through them
- You can also export your actions so that others can repeat them

Examples

- Export your project data and workflow to a file
  - Use ‘Export’ > ‘OpenRefine project archive to file’
  - This downloads a .tar.gz file to your downloads folder, which can be imported into OpenRefine if shared with others
- Careful – it will contain all data worked on, including old versions!

Pulling Data from an API

- API = Application Programming Interface
- APIs are ways you can access data from web databases directly
- OpenRefine can pull data from a range of sources
  - See Lesson on Advanced functions for an example for journal data
  - There's also an experimental example using VIAF (Virtual International Authority File – useful for famous named individuals
- OpenRefine can also import data from any web page, but may need extensive cleanup

Matching with Wikidata (reconciling)

- You can also match with WikiBase and DBPedia, database versions of Wikipedia
- This can match terms against articles in Wikipedia, and can import column data from there
- Wikipedia is semi-structured, so you'll need to know a bit about how your target Wikipedia articles are created

Recording Workflows for Reproducibility and Backups

Rationale

- Undo / redo is saved
- Exporting your steps as JSON
- Projects retain all your steps and undo data
- Tracking your history allows you to review whether steps have affected or distorted your data
- Allows others to verify your working

Exporting your workflow history

- In your project, choose 'export' then 'OpenRefine project archive to file'
- You can also save project history to Google Drive or Sheets
- This will contain all your history and all data, so be careful not to share sensitive work

Taking it Further

Documentation

- Read the documentation!
- You may also want to check the Stack Overflow OpenRefine tag or the OpenRefine Gitter room.
- There's a wide-ranging community of users

Trial and 'Error'

- Remember that OpenRefine is a safe way to play and experiment with data
- Remember to export your workflow/history every so often (e.g. at 'milestones')
- Explore new features and plugins once you are confident

Troubleshooting

OpenRefine only opens 'command box'

- 'Failed to bind to /127.0.0.1:3333'
  - You have a previous OpenRefine running – find the command-line window and close it down
  - Check whether any of your other applications is using :3333