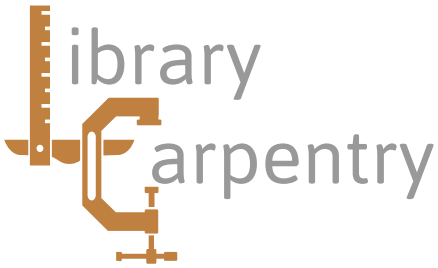

Software skills for librarians

Module 2: Open Refine
Exercises

1. a. Using the projects.tsv file create a new Open Refine project and call it 'Student projects exercise' or something similar.
   b. Do the same, but try changing some of the import options. How does this affect your data?

2. This time try importing the methods.xml file. You will need to select the XML tag which forms the basis for a single row of your data, the obvious one to choose in this case is the `<method>` tag. Don't worry too much about the contents unless you're a bellringer!

3. a. Using the student projects data, try to find all the projects which have 'video' in the title.
   b. What about 'audio'?
   c. Using a simple regular expression find all projects with XML or HTML in the title.

4. a. How would you find out which values have been used in the 'degree' column?
   b. Which is the most common course?
   c. Some records have a '2' in the degree column. Investigate these and delete them.

5. Create a numeric facet on the 'date' column. What do the histograms represent? Is there any long term trend?

6. Do the same, but first select only the Diploma projects from the 'degree' column. What can you infer from these results?

7. Remove all facets you have created, and once more create a text facet on the 'degree' column. Can you spot the deliberate inconsistencies? Now try ammending the values in the facet to correct these.

8. There are also some inconsistencies in the 'College' column. Explore these using a text facet, and then try to correct them using clustering. Now correct some of the remaining anomalies manually.

9. a. Create a new project using the crsids.tsv file.
   b. Using a transformation, correct any crsids which are include uppercase letters.
   c. Likewise, remove any trailing spaces from the crsid column.
   d. How would you have applied these transformations without using the menu?

10. The aim of this exercise is to prepare the crsids project so that we can use it to look up an identifier by name and add it to the projects file:

    a. Create a new column for the student's full name.
    b. Use the concatenation operator to join the surname and forename with a comma.
    c. Using a transformation, remove any trailing spaces from the crsid column.

11. This exercise will now add the crsids to the projects file:

    a. Switch back to the student projects and add a new column for the crsid.
    b. Using the cross function look up the crsid using the full name as a key.