Software skills for librarians

Module 2: Open Refine Answers

1a. This is simply a matter of selecting the ‘create project’ tab, setting various options including the character encoding, and the ‘parse cell text into numbers and dates’ option. Your data should look like this:

Name	College	Title	Year	Degree
Breslin, Catherine	Unknown	The multiple regression...	2004	MPhil
Buyanovsky, Olga	Unknown	Tree based state clustering...	2004	MPhil
Engstrom, Charlotta	Unknown	Topic dependence in sentiment...	2004	MPhil
Gibson, Matthew	Unknown	Efficient MLLR	2004	MPhil

1b. The obvious thing to change is the character encoding. The file uses only conventional ASCII characters, so most encodings will work, including UTF-8 and ISO 8859-1. UTF-16 treats each two adjacent characters as a single 16-bit value which results in the text appearing in Chinese ideograms. Expect similarly strange results if you try one of the Cyrrilic encodings, for example.

2. Again, this is simply a case of selecting the right options. If you choose the `<method>` tag as a basis for the data, the result will look something like this:

id	name	title	notation	classification
id0027	First	First Bob Doubles	3.1.125.1.125,1	Bob
id0028	Ancaster	Ancaster Bob Doubles	3.1.125.1.345,1	Bob
id0029	Quirister	Quirister Bob Doubles	3.1.125.1.5,123	Bob
id0030	Camelion	Camelion Bob Doubles	3.1.125.1.5,125	Bob

3a. Select filtering on the title column and enter `video`. There are a total of 29 projects.
3b. As before, but filter on `audio`; it retrieves 18 projects.
3c. Switch the filter to a regular expression and enter `(X|HT)ML`, or `XML|HTML` (the two are equivalent). This returns 4 projects.

4a. Create a text facet on the degree column. The most common course is the Tripos with nearly 900 projects.
4c. Create a text facet on the degree column and select the ‘2’ group. Delete them using the edit rows -> remove all matching rows option from the ‘all’ column. The errant entries are:

Name	College	Year	Degree
Parish, Tim	Fitzwilliam	0	2
Pilkington, Nicholas	Magdalene	0	2
Pinnis, Marcis	St. Edmunds	0	2
Raeesy, Zeynabalsadat	Lucy Cavendish	0	2

5. The number of projects, of all types, received each year. There is no long term trend, but a slight reduction in the middle.

6. There are very few diploma projects, and the course was withdrawn after 2008.

7. Some are labelled M.Phil CSTIT, it is reasonable to assume that these should be changed to M.Phil. Either select edit in the facet panel, or in one of the cells and then apply the change to all identical values.

8. Create a text facet on the college column and then select cluster. The default options do a reasonable job of correcting variants of the names of 11 colleges. This leaves three values with one instance each for variants of Gonville & Cauis, Pembroke and Trinity Hall which can easily be corrected manually. If you have time, there’s probably more you could do.

As an aside the previous two exercises demonstrate the value of using a code as an identifier, and then looking up the human-readable text as needed. For example, the projects database uses the letter ‘M’ internally for an M.Phil project. The inconsistences in this version of the data were deliberately added. On the other hand, the minor variations in the college names are all too real as the data originally came from many sources. I should have adopted codes like ‘JOHNS’ for St. John’s College, and used another database table to look these up.

9a. Create a new project and accept the default options. The data looks like:

Forenames	Surname	crsid
Immad	Akhund	IA236
Edward Mark	Allcutt	EMA29
Hussain	Almusaad	HA251
Alexander	Anderson	AA371

9b. Select edit cells -> common transforms -> to lowercase from the menu in the crsid column.
9c. On the same column enter `value.trim()`, or select it from the menu.
9d. Enter the transform: `value.toLowercase()`

10a. Create a new column based on the Surname column.
10b. Populate this column by entering the expression:
`join([cells.Surname.value, cells.Forenames.value], ", ")`, or
`cells.Surname.value+", "+cells.Forenames.value`

11. Add a new column based on the Name column, and enter the following expression:
`cell.cross("studentids", "fullname").cells.crsid.value[0]`.
Note that you may encounter a bug in Open Refine, in which case make sure that the project name doesn't contain any spaces, and, if necessary, restart Open Refine before trying this.

Closing remark

Some participants asked about the location of the project files. Open Refine creates these automatically, but doesn't obviously tell you the location. If you need to access files outside of Open Refine for whatever reason, then you'll find them at:

On Linux: `~/.local/share/openrefine/`
On MacOS: `~/Library/Application Support/OpenRefine/`
On Windows: `C:\Documents and Settings\user\Local Settings \Application Data\OpenRefine`
or `C:\Users\user\AppData\Roaming\OpenRefine`
or `C:\Users\user\OpenRefine`