Working with Data using OpenRefine

Owen Stephens
British Library

Using these slides

These slides were developed by Owen Stephens (owen@ostephens.com) on behalf of the British Library.

It is suggested when crediting this work, you include the phrase "Developed by Owen Stephens on behalf of the British Library"

Introductions

Outline for today

- Introductions and Outline
- Getting started
- Break (11:00)
- Basic OpenRefine functions
- Lunch (12:30-1:30)

Outline for today

- Bring your own data
- More OpenRefine
- Break (2:30-2:45)
- Review
- Finish (4:00)

“a tool for working with messy data”

http://openrefine.org

OpenRefine can help when…

- you have data in a simple tabular format
- there are inconsistencies in how the data is formatted
- there are inconsistencies in where data appears
- there are inconsistencies in terminology used in the data

OpenRefine can help you…

- Get an overview of a data set
- Resolve inconsistencies in a data set
- Help you split data up into more granular parts
- Match local data up to other data sets
- Enhance a data set with data from other sources

For example…

Data you have	Desired data
1st January 2014	2014-01-01
01/01/2014	2014-01-01
2014-01-01	2014-01-01
Jan 1 2014	2014-01-01

For example…

Data you have	Desired data
London	London
London]	London
London,]	London
london	London

For example…

Data you have	Desired data							
	Institution	Library name	Address 1	Address 2	Town/City	Region	Country	Postcode
University of Wales, Llyfrgell Thomas Parry Library, Llanbadarn Fawr, ABERYSTWYTH, Ceredigion, SY23 3AS, United Kingdom	University of Wales	Llyfrgell Thomas Parry Library	Llanbadarn Fawr		Aberystwyth	Ceredigion	United Kingdom	SY23 3AS
University of Aberdeen, Queen Mother Library, Meston Walk, ABERDEEN, AB24 3UE, United Kingdom	University of Abderdeen	Queen Mother Library	Meston Walk		Aberdeen		United Kingdom	AB24 3UE
University of Birmingham, Barnes Library, Medical School, Edgbaston, BIRMINGHAM, West Midlands, B15 2TT, United Kingdom	University of Birmingham	Barnes Library	Medical School	Edgbaston	Birmingham	West Midlands	United Kingdom	B15 2TT
University of Warwick, Library, Gibbett Hill Road, COVENTRY, CV4 7AL, United Kingdom	University of Warwick	Library	Gibbett Hill Road		Coventry		United Kingdom	CV4 7AL

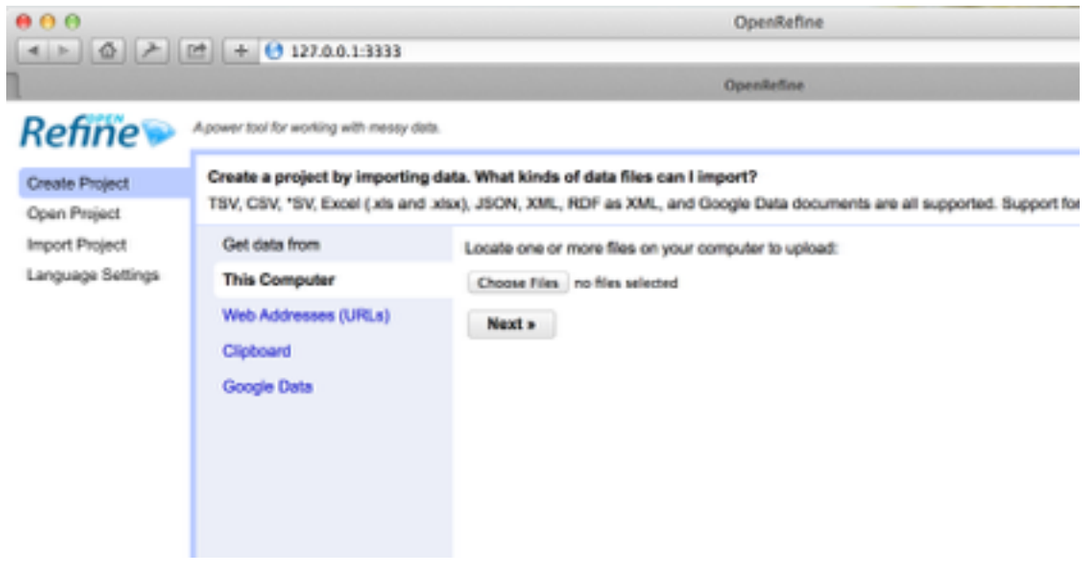
For example…

Data you have	Date of Birth from VIAF (Virtual International Authority File)	Date of Death from VIAF (Virtual International Authority File)
Braddon, M. E. (Mary Elizabeth)	1835	1915
Rossetti, William Michael	1829	1919
Prest, Thomas Peckett	1810	1879

Getting help

- The OpenRefine Wiki https://github.com/OpenRefine/OpenRefine/wiki
- The ‘Free your metadata’ site http://freeyourmetadata.org/ and book http://book.freeyourmetadata.org
- The OpenRefine mailing list and forum http://groups.google.com/d/forum/openrefine

Start using OpenRefine



OpenRefine vs Google Refine (I'll probably just say 'Refine')

Start Open/Google Refine on your laptops

Refine is a web application which runs locally on your PC. You access the interface through the browser

[switch to notes and walk through - get them using Refine at this point]

Hands-on!

[switch to notes and walk through - get them using Refine at this point]

Work up to Exercise 6

Transformations can help you

- Split data that is in a single column into multiple columns (e.g. splitting an address into multiple parts)
- Standardise the format of data in a column without changing the values (e.g. removing punctuation or standardising a date format)
- Extract a particular type of data from a longer text string (e.g. finding ISBNs in a bibliographic citation)

Example transformations

value.toUppercase()

‘value’ is always whatever is currently in the cell

value	value.toUppercase()
Hello	HELLO
Hello world	HELLO WORLD

Demonstrate the following in OpenRefine: toUppercase, toLowercase, toTitlecase, trim(), substring, replace, ‘+’ for concatenation

value.replace(search,replace)

Data types

- String
- Number
- Date
- Boolean
- Array

Arrays

A list of values

[“Monday”,”Tuesday”,”Wednesday”,”Thursday”,”Friday”,
”Saturday”,”Sunday”]

In OpenRefine arrays are usually created by transforming strings. For example the above array could be created through the following expression:

“Monday,Tuesday,Wednesday,Thursday,Friday,Saturday,
Sunday”.split(“,”)

Things you can do with an Array

- array[0]
- array.sort()
- array.uniques()
- array.join(“,”)

Getting data from elsewhere in your OpenRefine project

cells[“column name”].value

Hands-on!

Show Undo/Redo

Do Exercise 7

Regular Expressions

- A way of representing patterns in text strings
- “wildcards on steroids” (http://www.regular-expressions.info)
- Regular expressions let you:
  - Match on types of character (e.g. ‘upper case letters’, ‘digits’, ‘spaces’, etc.)
  - Match patterns that repeat any number of times
  - Capture the parts of the original string that match your pattern

Regular Expressions

/organise/

Character classes

[<list/range to be matched>]

[ABC]

[A-Z]

[123]

[0-9]

[A-Za-z0-9]

Regular Expressions

/organi[sz]e/

Character classes

.

\d

\w

\s

^

$

Regular Expressions

/^[Oo]rgani.e$/

Repetition

*

+

?

{min,max}

Regular Expressions

/.*/

/colou?r/

/\d{4}/

Capture groups

(capture this)

Regular Expressions

London : Mandarin, 1994

/.* : .*, \d{4}/

/(.*) : .*, \d{4}/

/.* : (.*), \d{4}/

/(.*) : (.*), (\d{4})/

GREL functions with Regular Expressions

- match()
- replace()
- split()

- match(string or regexp): Returns an array of the groups matching the given regular expression
- replace(string s, string or regex f, string r): Returns the string obtained by replacing f with r in s
- split(string s, string or regex sep, optional boolean preserveAllTokens): Returns the array of strings obtained by splitting s with separator sep. If preserveAllTokens is true, then empty segments are preserved.

Hands-on!

Do Exercise 8

Your data

- Type of data?
- Format?
- Size?
- What do you need to do?

Is OpenRefine the right tool?

- Excel
- Google Spreadsheets
- Google Fusion Tables
- Text editor
- Unix tools
- Writing code

Excel - familiarity, better for data entry, cut and paste operation, no paging to navigate

Google Spreadsheets - similar to Excel, can get external data relatively easily, easy to collaborate and share

Google Fusion Tables - if you just want to filter, easy to share

Text editor - powerful text editor can do many things

Unix tools - more challenging to use, but quick and some things (finding things, sorting) are easy

Writing code - most sophisticated and most to learn!

Advanced OpenRefine

- Retrieving data from online sources
- Using ‘Reconciliation’ services to match local data to external data sources
- Comparing data across two Refine projects
- Records and Rows

‘cross’

“Titles”:

ISBN	Title
9780141500348	Delilah Darling is in the Library
9781406305678	Library Lion

“Prices”:

ISBN	Price
9780141500348	6.99
9781406305678	5.99

Based on the ISBN column in the “Titles” project:

cell.cross(“Prices”,”ISBN”).cells[“Price]”.value[0]