Online intro to OpenRefine

– how to clean messy data

Marianne Gauffriau – mgau@kb.dk
Erik Schwägermann – es@kb.dk

Copenhagen University Library, The Royal Danish Library

UNIVERSITY OF COPENHAGEN

Agenda

Introduction to OpenRefine

Exercises

- Start OpenRefine
- Import csv file into OpenRefine
- Clean data in OpenRefine via graphical user interface (click)
- Clean data in OpenRefine via non-graphical user interface (run scripts)
- Export from OpenRefine to csv or excel

Take home messages and looking ahead

Learning objectives

Basic skills in how to use OpenRefine

Find out whether OpenRefine is useful in relation to our data

Knowledge about how to work with OpenRefine after the course

Introduction to OpenRefine

OpenRefine in context
- where OpenRefine is useful and where it's not

Collect data ☹

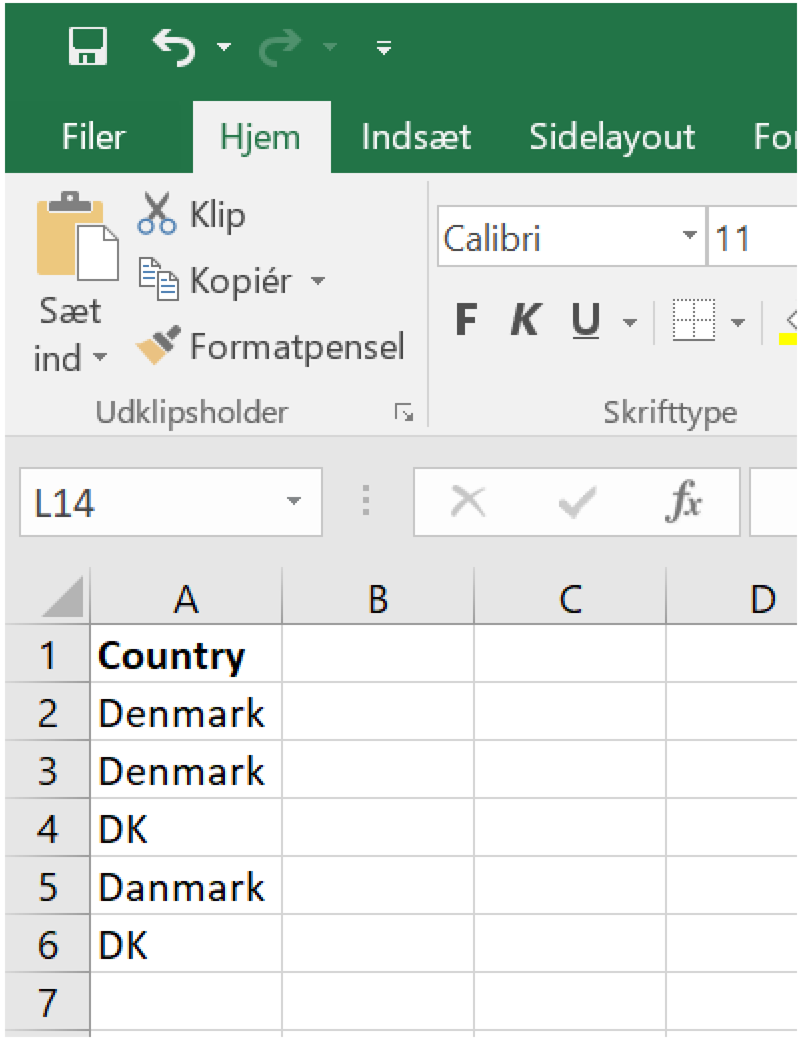
→ Import data into OpenRefine

Clean data ☺

Export data from OpenRefine →

Analyze data ☹

Visualize data ☹



	A	B	C	D
1	**Country**			
2	Denmark			
3	Denmark			
4	DK			
5	Danmark			
6	DK			
7				

OpenRefine – tabular data format

Type of data (comment)

- Long texts - meta data, coded data ☺

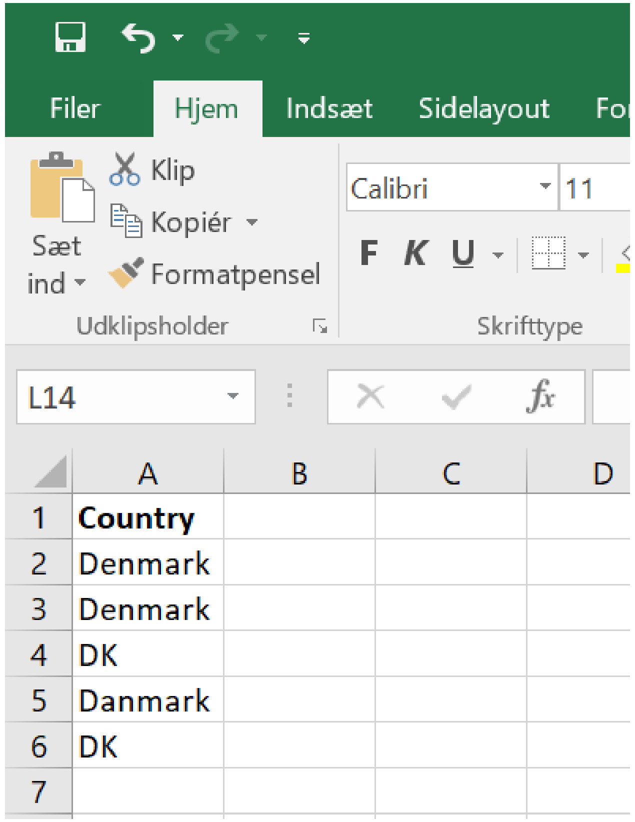
- Short texts ☺

- Images - meta data, coded data ☺

- Statistics - raw data ☺

- Surveys ☺

- …



L14

	A	B	C	D
1	**Country**			
2	Denmark			
3	Denmark			
4	DK			
5	Danmark			
6	DK			
7				

Exercises

Before we start

- OpenRefine (https://openrefine.org/download.html) is installed on you computer?
- Data file (https://ndownloader.figshare.com/files/11502815) is saved on your computer? Or you work on your own excel data file?
- Open exercises: https://datacarpentry.org/openrefine-socialsci/

Four exercises

1 Creating a new OpenRefine project - instructor and group

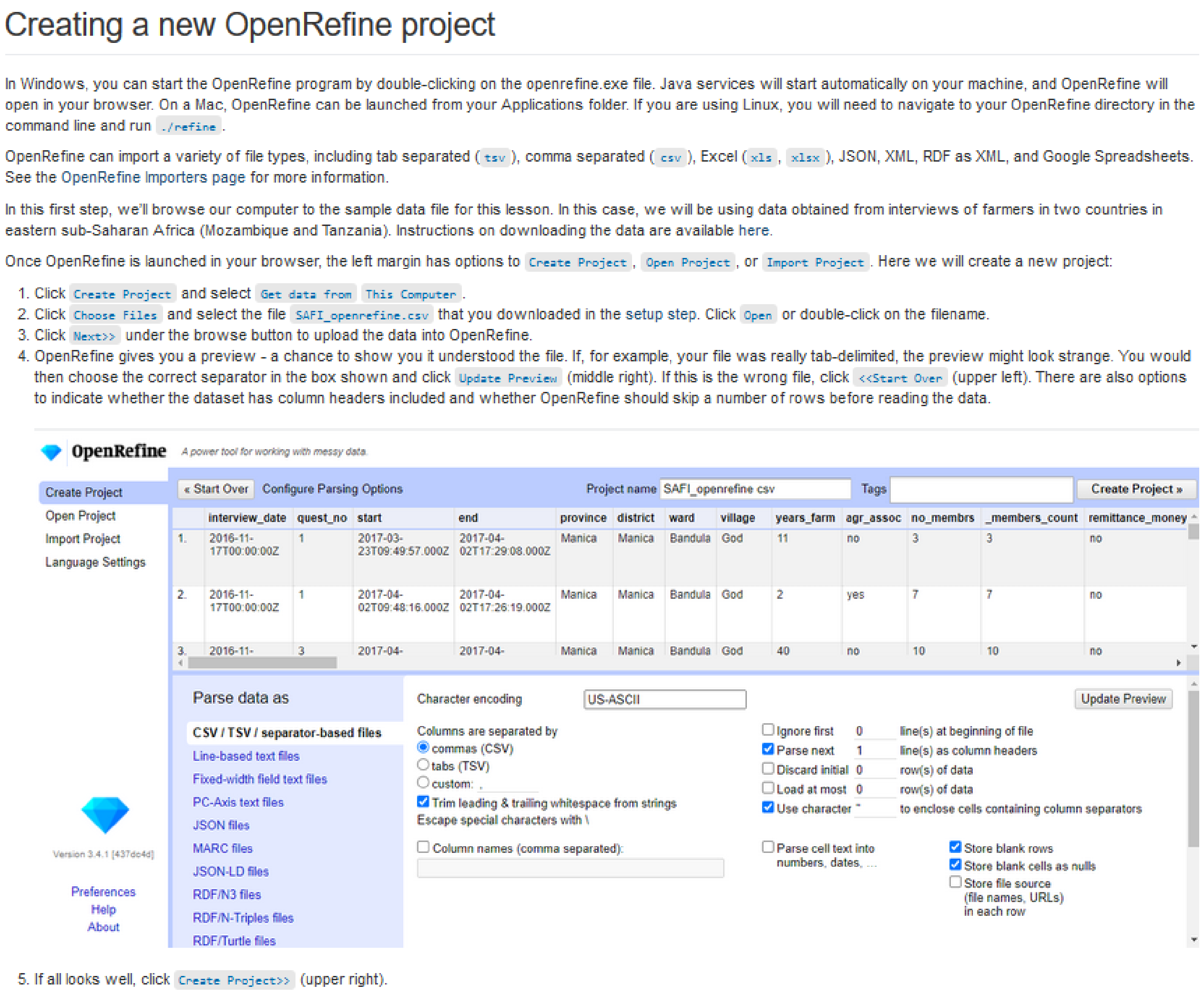
2 Using Facets - individual + discussion in plenum

3 Transforming data - individual + discussion in plenum

4 Exporting cleaned data - instructor and group

1 Creating a new OpenRefine project (1/3) – import csv file

https://datacarpentry.org/openrefine-socialsci/02-working-with-openrefine/index.html



Creating a new OpenRefine project

In Windows, you can start the OpenRefine program by double-clicking on the openrefine.exe file. Java services will start automatically on your machine, and OpenRefine will open in your browser. On a Mac, OpenRefine can be launched from your Applications folder. If you are using Linux, you will need to navigate to your OpenRefine directory in the command line and run `./refine`.

OpenRefine can import a variety of file types, including tab separated (`tsv`), comma separated (`csv`), Excel (`xls`, `xlsx`), JSON, XML, RDF as XML, and Google Spreadsheets. See the OpenRefine Importers page for more information.

In this first step, we'll browse our computer to the sample data file for this lesson. In this case, we will be using data obtained from interviews of farmers in two countries in eastern sub-Saharan Africa (Mozambique and Tanzania). Instructions on downloading the data are available here.

Once OpenRefine is launched in your browser, the left margin has options to `Create Project`, `Open Project`, or `Import Project`. Here we will create a new project:

1. Click `Create Project` and select `Get data from` `This Computer`.
2. Click `Choose Files` and select the file `SAFI_openrefine.csv` that you downloaded in the setup step. Click `Open` or double-click on the filename.
3. Click `Next>>` under the browse button to upload the data into OpenRefine.
4. OpenRefine gives you a preview - a chance to show you it understood the file. If, for example, your file was really tab-delimited, the preview might look strange. You would then choose the correct separator in the box shown and click `Update Preview` (middle right). If this is the wrong file, click `<<Start Over` (upper left). There are also options to indicate whether the dataset has column headers included and whether OpenRefine should skip a number of rows before reading the data.

Project name: SAFI_openrefine csv
Tags:

	interview_date	quest_no	start	end	province	district	ward	village	years_farm	agr_assoc	no_membrs	_members_count	remittance_money
1.	2016-11-17T00:00:00Z	1	2017-03-23T09:49:57.000Z	2017-04-02T17:29:08.000Z	Manica	Manica	Bandula	God	11	no	3	3	no
2.	2016-11-17T00:00:00Z	1	2017-04-02T09:48:16.000Z	2017-04-02T17:26:19.000Z	Manica	Manica	Bandula	God	2	yes	7	7	no
3.	2016-11-	3	2017-04-	2017-04-	Manica	Manica	Bandula	God	40	no	10	10	no

Parse data as
- **CSV / TSV / separator-based files**
- Line-based text files
- Fixed-width field text files
- PC-Axis text files
- JSON files
- MARC files
- JSON-LD files
- RDF/N3 files
- RDF/N-Triples files
- RDF/Turtle files

Character encoding: US-ASCII

Columns are separated by
● commas (CSV)
○ tabs (TSV)
○ custom: ,
☑ Trim leading & trailing whitespace from strings
Escape special characters with \
□ Column names (comma separated):

□ 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 most 0 row(s) of data
☑ Use character " to enclose cells containing column separators

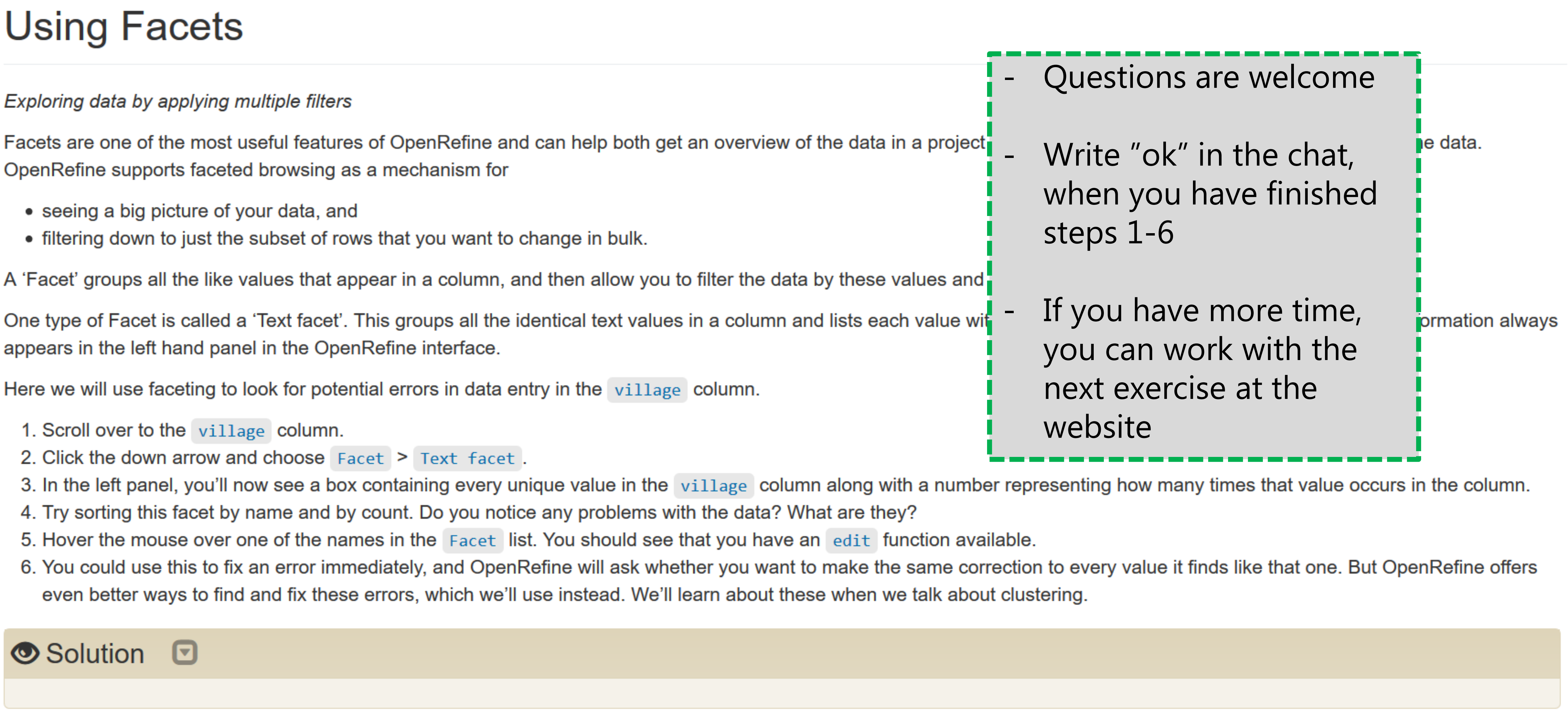
□ Parse cell text into numbers, dates, ...

☑ Store blank rows
☑ Store blank cells as nulls
□ Store file source (file names, URLs) in each row

5. If all looks well, click `Create Project>>` (upper right).

2 Using Facets (2/3) – get an overview of your data

https://datacarpentry.org/openrefine-socialsci/02-working-with-openrefine/index.html



Using Facets

Exploring data by applying multiple filters

Facets are one of the most useful features of OpenRefine and can help both get an overview of the data in a project e data. OpenRefine supports faceted browsing as a mechanism for

- seeing a big picture of your data, and
- filtering down to just the subset of rows that you want to change in bulk.

A 'Facet' groups all the like values that appear in a column, and then allow you to filter the data by these values and

One type of Facet is called a 'Text facet'. This groups all the identical text values in a column and lists each value wi ormation always appears in the left hand panel in the OpenRefine interface.

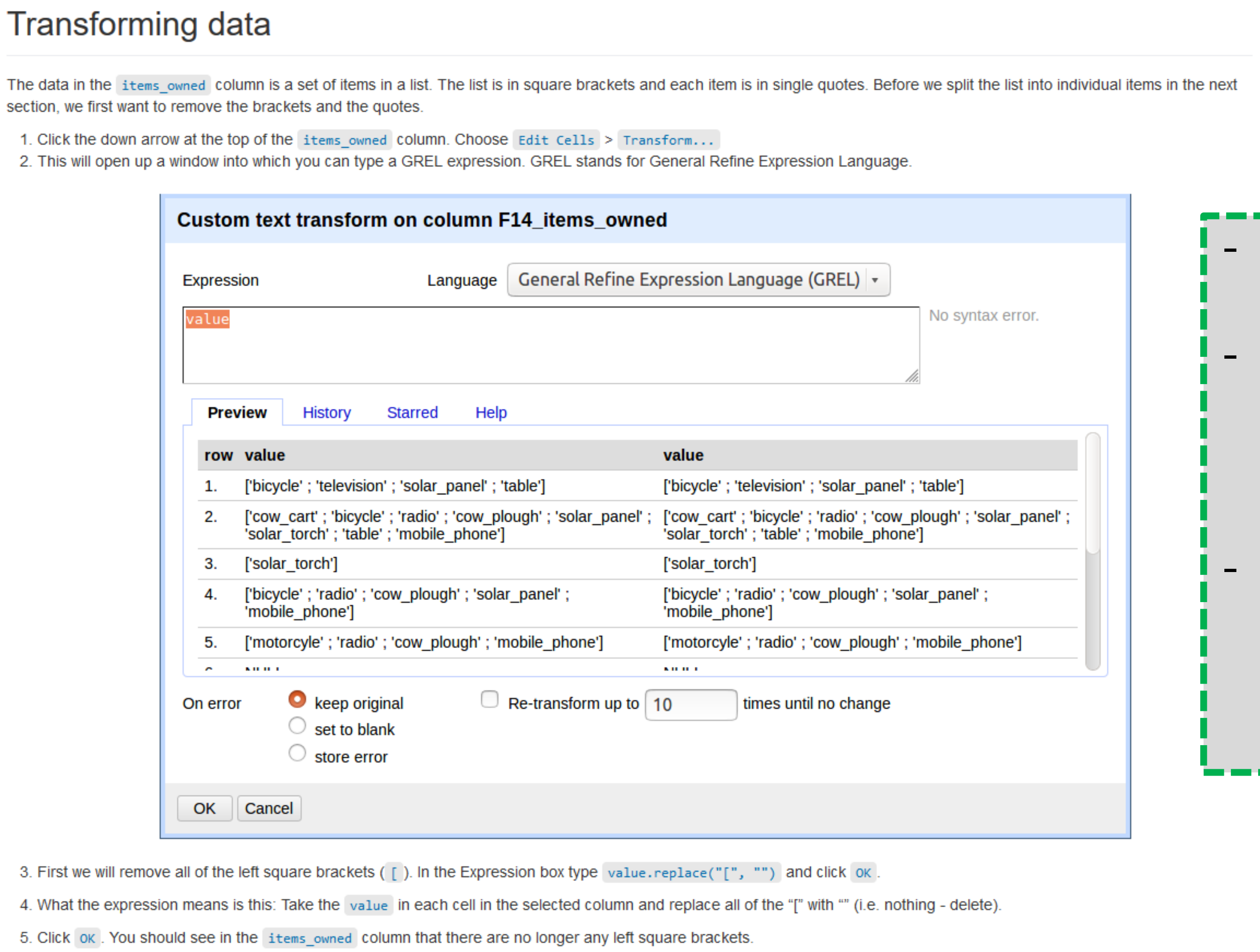
Here we will use faceting to look for potential errors in data entry in the `village` column.

1. Scroll over to the `village` column.
2. Click the down arrow and choose `Facet` > `Text facet`.
3. In the left panel, you'll now see a box containing every unique value in the `village` column along with a number representing how many times that value occurs in the column.
4. Try sorting this facet by name and by count. Do you notice any problems with the data? What are they?
5. Hover the mouse over one of the names in the `Facet` list. You should see that you have an `edit` function available.
6. You could use this to fix an error immediately, and OpenRefine will ask whether you want to make the same correction to every value it finds like that one. But OpenRefine offers even better ways to find and fix these errors, which we'll use instead. We'll learn about these when we talk about clustering.

- Questions are welcome
- Write "ok" in the chat, when you have finished steps 1-6
- If you have more time, you can work with the next exercise at the website

3 Transforming data (3/3) – clean data via GREL expressions

https://datacarpentry.org/openrefine-socialsci/02-working-with-openrefine/index.html



Transforming data

The data in the `items_owned` column is a set of items in a list. The list is in square brackets and each item is in single quotes. Before we split the list into individual items in the next section, we first want to remove the brackets and the quotes.

1. Click the down arrow at the top of the `items_owned` column. Choose `Edit Cells` > `Transform...`
2. This will open up a window into which you can type a GREL expression. GREL stands for General Refine Expression Language.

Custom text transform on column F14_items_owned

Expression

Language: General Refine Expression Language (GREL)

```
value
```

No syntax error.

row	value	value
1.	['bicycle' ; 'television' ; 'solar_panel' ; 'table']	['bicycle' ; 'television' ; 'solar_panel' ; 'table']
2.	['cow_cart' ; 'bicycle' ; 'radio' ; 'cow_plough' ; 'solar_panel' ; 'solar_torch' ; 'table' ; 'mobile_phone']	['cow_cart' ; 'bicycle' ; 'radio' ; 'cow_plough' ; 'solar_panel' ; 'solar_torch' ; 'table' ; 'mobile_phone']
3.	['solar_torch']	['solar_torch']
4.	['bicycle' ; 'radio' ; 'cow_plough' ; 'solar_panel' ; 'mobile_phone']	['bicycle' ; 'radio' ; 'cow_plough' ; 'solar_panel' ; 'mobile_phone']
5.	['motorcyle' ; 'radio' ; 'cow_plough' ; 'mobile_phone']	['motorcyle' ; 'radio' ; 'cow_plough' ; 'mobile_phone']

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

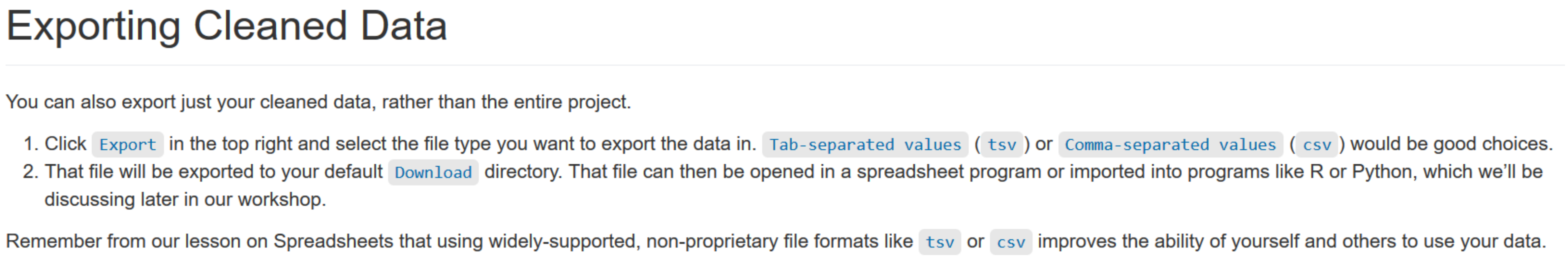
□ Re-transform up to 10 times until no change

3. First we will remove all of the left square brackets (`[`). In the Expression box type `value.replace("[", "")` and click `OK`.
4. What the expression means is this: Take the `value` in each cell in the selected column and replace all of the "[" with "" (i.e. nothing - delete).
5. Click `OK`. You should see in the `items_owned` column that there are no longer any left square brackets.

- Questions are welcome
- Write "ok" in the chat, when you have finished steps 1-5
- If you have more time, you can work with the next exercise at the website

4 Exporting Cleaned Data 1/1 – export from OpenRefine to csv

https://datacarpentry.org/openrefine-socialsci/06-saving/index.html



Exporting Cleaned Data

You can also export just your cleaned data, rather than the entire project.

1. Click `Export` in the top right and select the file type you want to export the data in. `Tab-separated values` (`tsv`) or `Comma-separated values` (`csv`) would be good choices.
2. That file will be exported to your default `Download` directory. That file can then be opened in a spreadsheet program or imported into programs like R or Python, which we'll be discussing later in our workshop.

Remember from our lesson on Spreadsheets that using widely-supported, non-proprietary file formats like `tsv` or `csv` improves the ability of yourself and others to use your data.

Take home messages and looking ahead

OpenRefine and your data?

Please, take 2 minutes to comment in the chat:

Is OpenRefine useful in relation to your data?

- Yes, because …
- No, because …

Purpose: Knowledge sharing in the group and feedback to the instructor.

Take home messages

OpenRefine

- Strengths: clean data + support many file-formats
- Weaknesses: data collection + data analysis + data visualization

Exercises

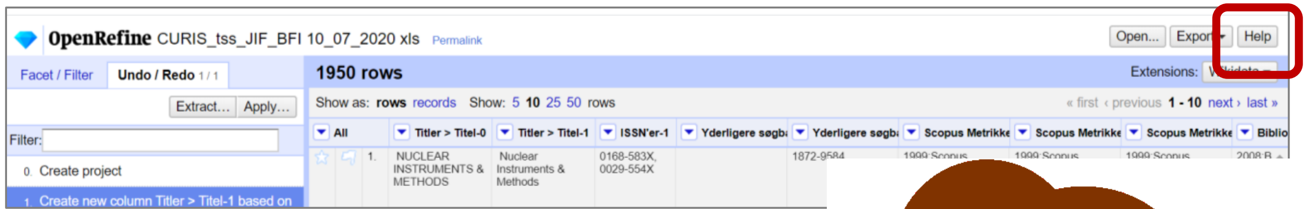
- OpenRefine via graphical user interface and non-graphical user interface
- Automatic documentation of expressions and easy reuse of expressions

Looking ahead

More exercises:

- https://datacarpentry.org/openrefine-socialsci/ and https://librarycarpentry.org/lc-open-refine/

Getting help: Use Help in OpenRefine, google it, or ask your librarian

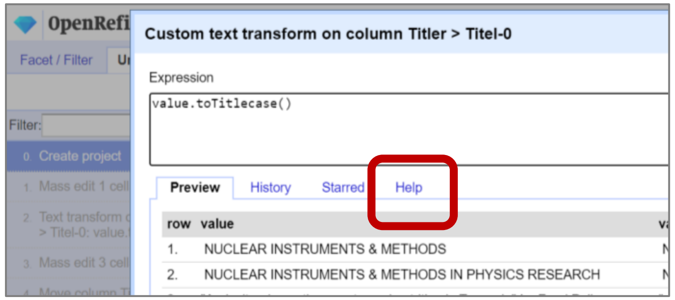


1950 rows

0. Create project

1. Create new column Titler > Titel-1 based on

	Titler > Titel-0	Titler > Titel-1	ISSN'er-1	Yderligere søgb	Yderligere søgb	Scopus Metrikke	Scopus Metrikke	Scopus Metrikke	Biblio
1.	NUCLEAR INSTRUMENTS & METHODS	Nuclear Instruments & Methods	0168-583X, 0029-554X		1872-9584	1999 Scopus	1999 Scopus	1999 Scopus	2008 B



Custom text transform on column Titler > Titel-0

Expression

```
value.toTitlecase()
```

0. Create project
1. Mass edit 1 cell
2. Text transform > Titel-0: value.
3. Mass edit 3 cell
4. Move column Ti

row	value
1.	NUCLEAR INSTRUMENTS & METHODS
2.	NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH

OpenRefine filter excel

Ask your librarian: kubdatalab@kb.dk