

# *Enqwyre*: User Guide

User-guide for the *Enqwyre* data-wrangling platform. Chapters in order of workflow.

[Overview of the Enqwyre approach to data wrangling](#)

[Create a new data wrangling process](#)

[Merge and/or assign a reference column for the spreadsheets](#)

[Structure the data to conform to the destination schema](#)

[Order](#)

[Order by \[newest / oldest\]](#)

[Join](#)

[Calculate](#)

[Categorise](#)

[Nested Methods](#)

[Categorise unique terms to conform to the schema categories](#)

[Transform and filter source data into destination schema](#)

[Error correction, and resubmission](#)

## Overview of the *Enqwyre* approach to data wrangling

***Enqwyre*'s objective is to offer a straightforward and rapid method for restructuring messy data to conform to a standardised metadata schema.**


Importing messy data into a single schema is a slow and tedious process, but needs to be done by a person with good technical skills and an intimate understanding of the data.

The wrangler's challenge is in thinking through how to restructure the data to conform to the required schema, but that involves repetitive cutting and pasting of columns, as well as writing simple arithmetic functions to either perform simple calculations, or join columns together.

While there are specialist data wrangling tools (e.g. OpenRefine or Trifacta), they are aimed at a full sequence of restructuring large datasets, as well as post-processing data, such as doing complex calculations, categorisations, and filtering.

The reality is that the most complex and time-consuming part is the restructuring. Once in an appropriate standardised format, a software system can take care of everything.

Most steps start from the main ***Process*** landing page.

Sqwyre

AboutPricingProcessAdmin [super]Logout




























Search

[Home](#) / [Process](#)

### Process Management

The current quarter began on 2019-07-01 and this is cycle 012.

Authorities Create

Authority ↑	Status	Last received	Last requested	Publication	Frequency	Links
> <a href="#">Adur</a>		2019-08-23	2019-08-23 	Self-published ▼	Monthly ▼	 
> <a href="#">Allerdale</a>		2019-08-23	2019-08-23 	Self-published ▼	▼	 
> <a href="#">Amber Valley</a>		2018-06-20	2018-04-06 	FOI ▼	▼	 
> <a href="#">Arun</a>		2018-04-07	2018-04-06 	FOI ▼	▼	 
> <a href="#">Ashfield</a>		2019-08-27	2019-08-23 	Self-published ▼	▼	 
> <a href="#">Ashford</a>		2018-11-06	2018-04-06 	FOI ▼	▼	 
> <a href="#">Aylesbury Vale</a>		2018-07-05		Self-published ▼	▼	 
> <a href="#">Babergh</a>		2018-06-07	2017-09-10 	Self-published ▼	▼	 
> <a href="#">Barking and Dagenham</a>		2018-05-06	2017-12-07 	Self-published ▼	▼	 

## Create a new data wrangling process


From the **Create** tab on the main **Process** landing page:

**Process Management**

The current quarter began on 2019-07-01 and this is cycle 012.

---

[Authorities](#)   [Create](#)

  
Drop your files here or click to upload

Cardiff Business Rates - Live Properties - Update 2 2018.xlsx

Cardiff Business Rates - Current Empty Exemptions - Update 2 2018.xlsx

Cardiff Business Rates - Current Empty Reliefs - Update 2 2018.xlsx

Cardiff Business Rates - Mandatory or Discretionary Rate Relief - Update 2 2018.xlsx

**Drag and drop** and spreadsheets making up the dataset for this process. **Search** for the name of the local authority, and assign a **cycle** and data **type** (*All* - for both occupied and vacant, or *Occupied* or *Vacant*, accordingly).

You can pick **any cycle** from the latest, back to the first. You will get a warning if you pick a cycle other than the current one, but you will not be stopped from uploading.

Once uploaded, the app will return to the **Process** landing page. Refresh the page to see any **Status** update changes. There are four main states:

1. Review **Merge**
2. Review **Structure**
3. Review **Categories**
4. Review **Transform**

**Check:** Each file must have a single header row, starting at the top-most, left-most cell of the sheet. Check that the bottom of the spreadsheet doesn't contain any weird artifacts.

**Warning:** you will **overwrite** any earlier saved dataset if you upload a new version.

**Error messages:**

- "Failed dependency" - check the files, it may be that they can't be opened.

## Merge and/or assign a reference column for the spreadsheets

Starting from the **Process** page, click on **Review Merge** and start the merge process.

> Cardiff Review Merge 2017-09-20 Self-published Quarterly

Merging serves two purposes:

- If more than one source, **combine the different spreadsheets** into a single merged file;
- To **identify the reference column** used to merge the source data into the database.

### Merge 'Cardiff' - Process Cycle 012

Merge / Structure / Categorise / Transform

Select the **billing reference** as the merge column for each file. Sort the files in order of merge. Select the reference column **even** if there is only one file.

Submit

Merge on

Valuation Office Ref

Source: Cardiff Business Rates - Live Properties - Update 2 2018.xlsx

	Full Property Address_x	Post Code_x	Current Rateable Value_x	Current Analysis Code	Valuation Office Ref	Primary Liabe party name_x	Account start_x	Primary Liabe party name_y	Full Property Address_y	Post Code_y	Current Property Exemption Code	Current Prop Exemption Start Date	Current Rateable Value_y	Primary Liabe party name_x1	Full Property Address_x1	Post Code_x1	Current Relief Type_x	Current Relief Award Start Date_x	Current Rateable Value_x1	Primary Liabe party name_y1	Full Prop Add
0	1st Fir At, 22, Princes Street, Roath, Cardiff...	CF24 3PS	690	CO	3211001540002222	Hafod Care Association Limited	2016-12-09					NaT						NaT		Hafod Care Association Limited	1st Fir
1	1st Fir Front, 5, High Street, Cardiff, CF10 1AW	CF10 1AW	7500	CO	321010000000005966	Personal data	2018-05-01	Personal data	1st Fir Front, 5, High Street, Cardiff, CF10 1AW	CF10 1AW	LIST BUILD	2018-05-01	7500.0					NaT			
2	S Barrack Lane, Castle, Cardiff, CF10 2FR	CF10 2FR	13750	CS	32101000041000522	Linc Cymru Housing Association Ltd	2018-10-31					NaT		Linc Cymru Housing Association Ltd	S Barrack Lane, Castle, Cardiff, CF10 2FR	CF10 2FR	EPRN	2018-12-25	13750.0		

Merge on

Valuation Office Ref

Source: Cardiff Business Rates - Current Empty Exemptions - Update 2 2018.xlsx

	Primary Liabe party name	Full Property Address	Post Code	Current Property Exemption Code	Current Prop Exemption Start Date	Valuation Office Ref	Current Rateable Value
0	A F Blakemore & Son Ltd T/A Spar	2, Station Road, Radyr, Cardiff, CF15 BAA	CF15 BAA	VOID	2018-10-15	323500001200000215	21750
1	A G Quidnet Uk Industrial 2 Bv	Ams Marine & Industrial Engineering Ltd, Freem...	CF11 8EQ	VOID 6	2018-09-03	322290001758000822	23250
2	A G Quidnet Uk Industrial 2 Bv	Unit 15c, Freemans Parc, Penarth Road, Granget...	CF11 8EQ	VOID 6	2018-09-03	322290001758001544	7800

Merge on

Valuation Office Ref

Source: Cardiff Business Rates - Current Empty Reliefs - Update 2 2018.xlsx

	Primary Liabe party name	Full Property Address	Post Code	Current Relief Type	Current Relief Award Start Date	Valuation Office Ref	Current Rateable Value
0	180 Wellness Centres Llp	2-3, Cleeve House, Lambourne Crescent, Cardiff...	CF14 5GP	EPRN	2017-11-21	321080001390099033	36000
1	3d Property Investments Ltd	7, St Andrews Crescent, Cardiff, CF10 3DA	CF10 3DA	EPRN	2018-04-15	3210100017200007AA	26750
2	3d Property Investments Ltd	Unit 1c & 1d, 14/15, Curran Road, Butetown, Ca...	CF10 5NE	EPRN	2015-09-05	321020000540001466	16750

Only the first few rows of each uploaded spreadsheet are shown. **Drag 'n drop** these into order.

Files are merged into the first/top file in order. Where there are common columns (e.g. *Postcode*) each additional column gets a suffix appended, starting with *\_x*, *\_y*, *\_z* (e.g. *Postcode*, *Postcode\_x*, *Postcode\_y*, etc.).

**Check:** Select a merge column for each file, even if there is only one file present.

## Structure the data to conform to the destination schema

Starting from the **Process** page, click on **Review Structure** and start the structure process.

> Cardiff      Review Structure      2017-09-20      Self-published      Quarterly

This is the critical wrangling process.

### Structure 'Caerphilly' - Process Cycle 012

Merge / Structure / Categorise / Transform

Drag 'n drop to build your methods for each column in the destination schema. Each column's method must start with a single action. You can create nested methods, and each must also start with a single action.

Submit

< > ● ○ ○ ○ ○ ○ ○ ○

+

-

≡

Billing Reference ★ ●

Order ×

ba\_ref 📄 ×

New

Order

Order by newest

Order by oldest

Calculate

Categorise

Join

	la_code	ba_ref	prop_empty	prop_empty_date	prop_occupied	prop_occupied_date	prop_ba_rates	occupant_name	postcode
0	W06000018	4400M0010070080300	N		Y		4000		CF83 3AT
1	W06000018	4400Z0010173064400	N		Y		860		NP11 6BW
2	W06000018	4400A0010060011100	N		Y		13500	AEL Y BRYN COMMUNITY CENTRE	NP22 5DR
3	W06000018	4400A0010115000300	N		Y		166000	BIFFA WASTE SERVICES LTD	CF48 4AB
4	W06000018	4400A0010115002800	N		Y		59500	BIFFA WASTE SERVICES LTD	CF48 3DB

la\_code 📄

ba\_ref 📄

prop\_empty 📄

prop\_empty\_date 📄

prop\_occupied 📄

prop\_occupied\_date 📄

prop\_ba\_rates 📄

occupant\_name 📄

postcode 📄

The page is divided into a number of sections:

1. **Table:** the first 5 rows of the merged spreadsheet;
2. **Actions:** **red buttons** indicating the wrangling actions to be performed;
3. **Fields:** **green buttons** indicating the fields in the table which can be selected for wrangling actions;
4. **Workspace:** the main work area where the destination schema is presented, and **methods** are defined;

A **Method** consists of a list of terms which starts with an **Action** and then consists of a series of **Fields**.

A method may contain additional nested methods, each returning a new field. Each method contains one, and only one, action. Methods can contain unlimited fields.

Actions may require that fields are defined by **Modifiers**.

The **Schema** (destination fields) are listed in the workspace. Traverse the terms with the left-right buttons. Completed terms change from **red** to **green**:



**Modifiers** are special terms used in specific actions:



Used in **Order by**, **Calculate**, and **Categorise** actions with a defined meaning in each.



Used in **Order by**, **Calculate**, and **Categorise** actions with a defined meaning in each.



Creates a new **Nested Method** and can be used in any action.

Each **Method** is created by dragging **Actions** and **Fields** into the Workspace, then dragging them into the correct order. **Modifiers** are added to the **Workspace** by clicking on them. They are added to the bottom of the current **Method**.

**Schema** fields, indicated with a coloured star, **Red** is compulsory, **Grey** is optional.

Billing Reference ★ ● Actual Rates Paid ★ ●

To **remove** an optional method, ensure no fields or actions are present in the workspace.

Rates Reliefs Categories ★ ●

**Actions** are defined as follows:

## Order

Merge columns, replacing blanks with subsequent columns in order. Combine a list of fields into a single method. Each field will be evaluated in order from top to bottom. The original term will be kept, unless it is blank.

Field 1	Field 2		Schema
a	g	➡	a
	k		k
b	l		b
c			c
d	m		d

Ratepayer Name ★ ●

Order ×

Primary Liable party name\_x 📅 ×

Primary Liable party name\_y 📅 ×


Primary Liable party name\_x.1 📅 ×

Primary Liable party name\_y.1 📅 ×


## Order by [newest / oldest]

Instead of selecting terms by field order, we can select terms by the date in which they were added. Terms can be selected either as **newest** or **oldest**.

The **newest** term is derived from a corresponding date field, for the most recent date:

Field 1	Date 1	Field 2	Date 2		Schema
a	2018-12-02	g	2018-10-02		a
	2018-01-14	k	2017-04-20		
b	2017-03-12	l	2018-03-23		l
c	2018-06-02		2017-04-04		c
d	2012-11-10	m	2015-12-12		m

The **oldest** term is derived from a corresponding date field, for the oldest date:

Field 1	Date 1	Field 2	Date 2		Schema
a	2018-12-02	g	2018-10-02		g
	2018-01-14	k	2017-04-20		k
b	2017-03-12	l	2018-03-23		b
c	2018-06-02		2017-04-04		
d	2012-11-10	m	2015-12-12		d

The **method** term field and date field combination is defined as: **<field> <+> <field>** where the **<+>** is the **modifier**. This creates a list as follows:

Occupation State Date ★ ●

Order by newest ×

Account Start date\_x 📅 ×

+ ×

Account Start date\_x 📅 ×

Current Prop Exemption Start Date\_x 📅 ×

+ ×

Current Prop Exemption Start Date\_x 📅 ×

Current Relief Award Start Date\_x 📅 ×

+ ×

Current Relief Award Start Date\_x 📅 ×

The order itself doesn't matter as the terms will be evaluated in date order.

**Note:** Where the fields themselves are dates, **both** the selection field and date field are the **same field name**.

## Join

Used to join text terms into a single phrase. For example, assume field 1 has a term “Justice”, and field 2 has a term “Department”. Joining them results in “Justice Department”.

It looks identical to Order, but the terms will be joined together into a single term. A space will be placed between them.

Ratepayer Name ★ ●

Join ×

Primary Liable party name\_x 📅 ×

Primary Liable party name\_y 📅 ×

## Calculate

Calculations are only simple additions or subtractions. More complex analysis will need to wait on import into the database.

Each **Field** imported into the **Method** must start with a **Modifier**, either + or - defining whether the term is positive or negative. All terms are summed, subject to their modifier.

Actual Rates Paid ★ ●

Calculate ×

+ ×

Current Rateable Value\_x 📅 ×

- ×

Current Rateable Value\_y 📅 ×

**Note**, it doesn't matter whether numbers in the columns are positive or negative. Only the modifier matters.

Not all numeric data requires a calculation. You can also **Order** numeric data.



## Categorise

There are two choices to be made for any field:

1. Does the column contain individual terms, each expressing **unique information**, that needs to be classified?
2. Are the **terms themselves unimportant**, and only the presence or absence of any term implies a boolean True or False belonging to a category?

As example:

Field 1	Field 2
dog	27
cat	
frog	rabbit
dog	
cat	2012-11-10

In Field 1, we want a list of the unique terms and will assign these to the predetermined categories from the schema (e.g. *dog* and *cat* are *mammals*, but *frog* is an *amphibian*).

In Field 2, we are not interested in the terms themselves. The presence of anything is *True*, and a blank field is *False*. So *27*, *rabbit* and *2012-11-10* are *True*.

We define this with the **modifiers**. Each action is defined as: **<modifier> <field>**



Identify **unique terms** for the column defined by this **field**.



Ignore specific terms. The presence of any term is regarded as **True**. Blanks are **False**.

The next step in the process, **Categorise**, is where you will assign the terms identified in this step to specific categories. These will then be added to the **Schema** destination field.

Occupation State ★ ●

Categorise x

+ x

Current Property Exemption Code\_x 📅 x

+ x

Current Relief Type\_x 📅 x

There is no order here. All the terms identified (whether modified by + or -) will be offered in the next step.

## Nested Methods

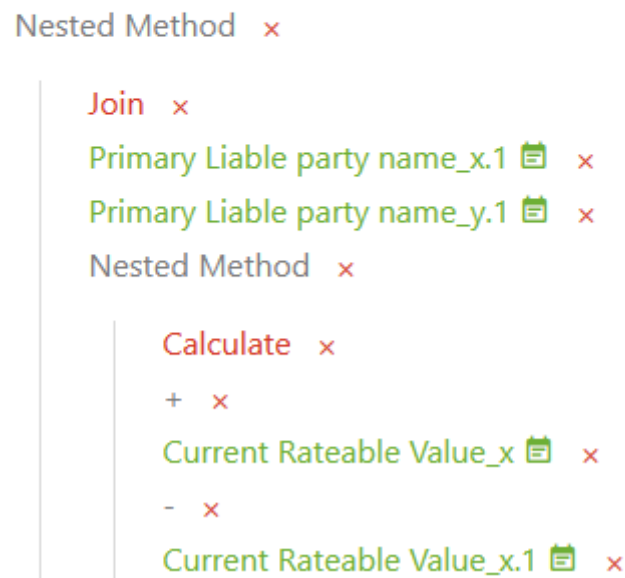


Creates a new **Nested Method** and can be used in any action.

Sometimes there need to be interim steps before you can complete method. Maybe you first need to join two sets of fields, then order the result:



Maybe you need to do a series of calculations, then order those results. And **Nested Methods** can contain **Nested Methods**.



In this way, you can produce quite complex methods to transform messy data.

**Note:** once all **Schema** methods are approved (green), click **Submit** to save.

## Categorise unique terms to conform to the schema categories

Starting from the **Process** page, click on **Review Categories** and start the categorisation process.

> Cardiff Review Categories 2017-09-20 Self-published Quarterly

The categorisation page offers a simple view:

**Categorise 'Cardiff' - Process Cycle 012**

[Merge](#) / [Structure](#) / [Categorise](#) / [Transform](#)

Drag 'n drop to build the terms in each destination schema to link them to the available categories. Submit

< > ⊙ ○

Occupation State

TRUE

FALSE

LIST BUILD

LOW RV

EXEMPT

VOID

INSOLVENT

LOW RV CAR

AD RIGHT

CAR SPACE

VOID 6

EPRN

EPRI

EPCA

EPCH

MAND

TDIS

DISC

CHSH

SMAN

STDI

There are two main areas:

1. **Source unique terms** on the right, listing all the unique terms identified from the **Fields** chosen in the **Structure** process;
2. **Workspace**, containing each of the **destination terms** for the **Schema**.

You can scroll through the destination schema category fields with the arrow buttons:



Simply **drag 'n drop** the **source terms** into the appropriate choices for the **destination terms**.

Occupation State

TRUE

FALSE

INSOLVENT x

VOID x

VOID 6 x

EPRN x

EPRI x

EPCA x

EPCH x

**Note:** once **Schema** categories are complete, click **Submit** to save.

## Transform and filter source data into destination schema

Starting from the **Process** page, click on **Review Transform** and start the transformation process.

> Cardiff Review Transform 2017-09-20 Self-published Quarterly

This page presents a very simple set of choices:

### Transform 'Cardiff' - Process Cycle 012

Merge / Structure / Categorise / Transform

Do you wish to import data after a specific date, or import all the data?

Submit

☐ All data ☐ Latest only ☒ After a specific date

Select a date

<

September

>

2019

>

Su	M	Tu	W	Th	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	1	2	3	4	5

There are three choices:

- **All data:** import all data in the file;
- **Latest only:** import only the latest term, sorted by the **Schema date** column, for the **Schema reference field** identified in the **Merge** step;
- **After a specific date:** pick a specific date from the calendar dialogue; if this is not the first import, then the **last import latest date will be displayed** and you could import from that date.

**Note:** no matter what you select, any duplicates in the data will automatically be removed, leaving only unique rows defined by the methods you create, and wrangled into the defined schema.

Once complete, **Submit** and begin the transformation process. All going well, this is the last step. If there are any issues, you will have an additional opportunity to correct the data.

> Cardiff Complete 2019-09-15 Self-published Quarterly

## Error correction, and resubmission

*Enqwyre* is designed to pick up and correct the majority of errors in the data while doing as little as possible to change that data format or structure. As long as the data can be validated against the schema, everything should be fine.

> Cardiff Import Error 2019-09-15 Self-published Quarterly

However, if the **Transform** process runs into an error, you can view the error page and **download** an interim process file. This will be a file that conforms to the current schema.

### Errors: 'Cardiff' - Process Cycle 012

[Merge](#) / [Structure](#) / [Categorise](#) / [Transform](#)

[Errors](#) Re-upload Data

**Download** the restructured data file and correct any errors. Don't change the file structure. **Re-upload** the corrected file when complete.

Download

You can then manually review the file.

If the program can identify any errors, you will see a table showing the rows - and row numbers - indicating what you should look for and correct.

[Errors](#) Re-upload Data

**Download** the restructured data file and correct any errors. Don't change the file structure. **Re-upload** the corrected file when complete.

Download

#### Spreadsheet row errors:

	ba_ref	prop_ba_rates	occupant_name	postcode	occupation_state	occupation_state_date	occupation_state_reliefs	la_code
13428	Pikhaya Error: missing compulsory term	112000	Greenwich Leisure Ltd	CF242SJ	True	2018-09-15 01:00:00+01:00	['exempt']	W06000015
13429	Pikhaya Error: missing compulsory term	112000	Greenwich Leisure Ltd	CF242SJ	True	2018-08-20 01:00:00+01:00	['exempt']	W06000015
13430	Pikhaya Error: missing compulsory term	112000	Greenwich Leisure Ltd	CF242SJ	True	2018-09-15 01:00:00+01:00	['charity']	W06000015
13431	Pikhaya Error: missing compulsory term	112000	Greenwich Leisure Ltd	CF242SJ	True	2018-08-20 01:00:00+01:00	['charity']	W06000015
13432	Pikhaya Error: missing compulsory term	24250	Cardiff Cc Education	CF242LZ	True	2018-09-15 01:00:00+01:00	[]	W06000015
13433	Pikhaya Error: missing compulsory term	24250	Cardiff Cc Education	CF242LZ	True	2018-08-20 01:00:00+01:00	[]	W06000015
13434	Pikhaya Error: missing compulsory term	24250	Cardiff Cc Education		True	2018-09-15 01:00:00+01:00	[]	W06000015
13435	Pikhaya Error: missing compulsory term	24250	Cardiff Cc Education		True	2018-08-20 01:00:00+01:00	[]	W06000015
13436	Pikhaya Error: missing compulsory term	24250	Cardiff Cc Education		True	2018-09-15 01:00:00+01:00	[]	W06000015
13437	Pikhaya Error: missing compulsory term	24250	Cardiff Cc Education		True	2018-08-20 01:00:00+01:00	[]	W06000015
13438	Pikhaya Error: missing compulsory term	24250	Cardiff Cc Education		True	2018-09-15 01:00:00+01:00	[]	W06000015
13439	Pikhaya Error: missing compulsory term	24250	Cardiff Cc Education		True	2018-08-20 01:00:00+01:00	[]	W06000015

On the same page, you will see the **Re-upload Data** tab:

[Errors](#) [Re-upload Data](#)



Drop your files here or click to upload

Submit

**Don't change the filename or file-type** when you edit the file, or it will not be accepted.

Should you wish to review previous methods, or make changes as new information comes in, you can check the **Status History** tab from the **Process** page for each item.

▼

Cardiff

Complete

2019-09-15

Self-published

▼

Quarterly

▼

Notes	Add Note	Status History	Update History			
Cycle	Status	Last process	Last received	Last requested	Publication	Links
012	Complete		2019-09-15		Self-published	
006			2017-09-20		Self-published	
005			2017-09-20		Self-published	
004			2017-09-20		Self-published	
003					Self-published	
002			2016-07-05	2016-10-17	Self-published	
001			2016-07-05		Self-published	

1

▼

Cheshire West and Chester

Complete

2018-02-22

Self-published

▼

▼

> Cheshire West and Chester

Complete

2018-02-22

Self-published ▼

▼

This completes the wrangling process.