OpenRefine

OpenRefine is a great tool for handling messy data. It is easy to install and the greatest benefit I found is security since it runs on our machines and doesn't require documents to be uploaded on the web. It is free and open source tool. It saves time.

Possible uses are to:

1] Handle messy data: Creating a facet provides the chance to view the different types of values and clustering them most of the time shows similar values but having variant cases or spellings, which could be corrected easily.

2] Easily transform data: It can easily transform data since there are many in-built transformations like to Lower available and we can create custom transformations as well.

3] Multiple language support: It provides support by providing General Refine Expression Language, Jython, Python and Clojure for custom transformations.

4] Revert the changes – It is easy to revert the changes and go back to any changed step since the changes made are available in the Undo/Redo section.

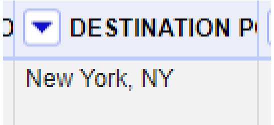
Example:

Let us take an example to demonstrate the use of OpenRefine. The dataset NJShipwreks.csv attached below was not a tidy dataset because of the following reasons.

NJShipwrecks.csv

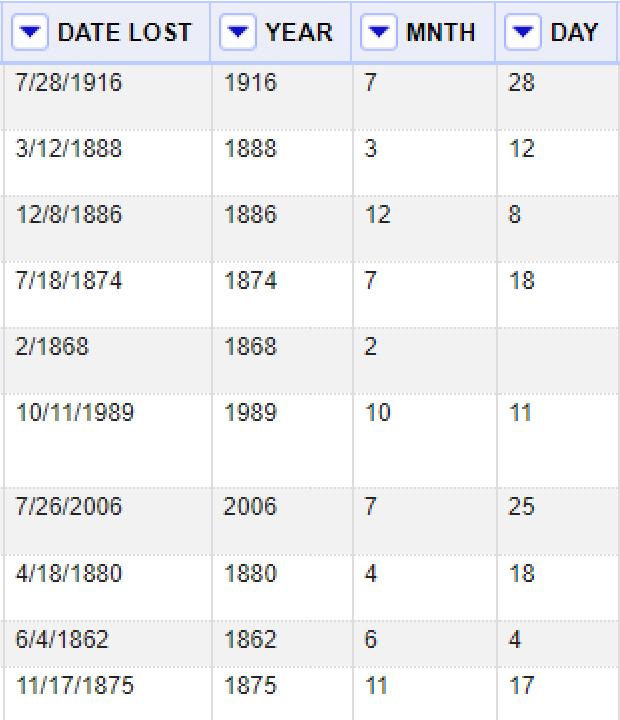
i] Multiple variables are stored in one column

In this dataset each variable doesn't form a column, e.g., in Departure Point city and state are clubbed together. They can be split into multiple columns.



DESTINATION P
New York, NY

Date Lost already has the date but there are again columns which state year month and day so the date lost column can be removed:



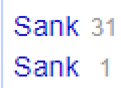
DATE LOST	YEAR	MNTH	DAY
7/28/1916	1916	7	28
3/12/1888	1888	3	12
12/8/1886	1886	12	8
7/18/1874	1874	7	18
2/1868	1868	2	
10/11/1989	1989	10	11
7/26/2006	2006	7	25
4/18/1880	1880	4	18
6/4/1862	1862	6	4
11/17/1875	1875	11	17

ii] Multiple types of observational units are stored in the same table.

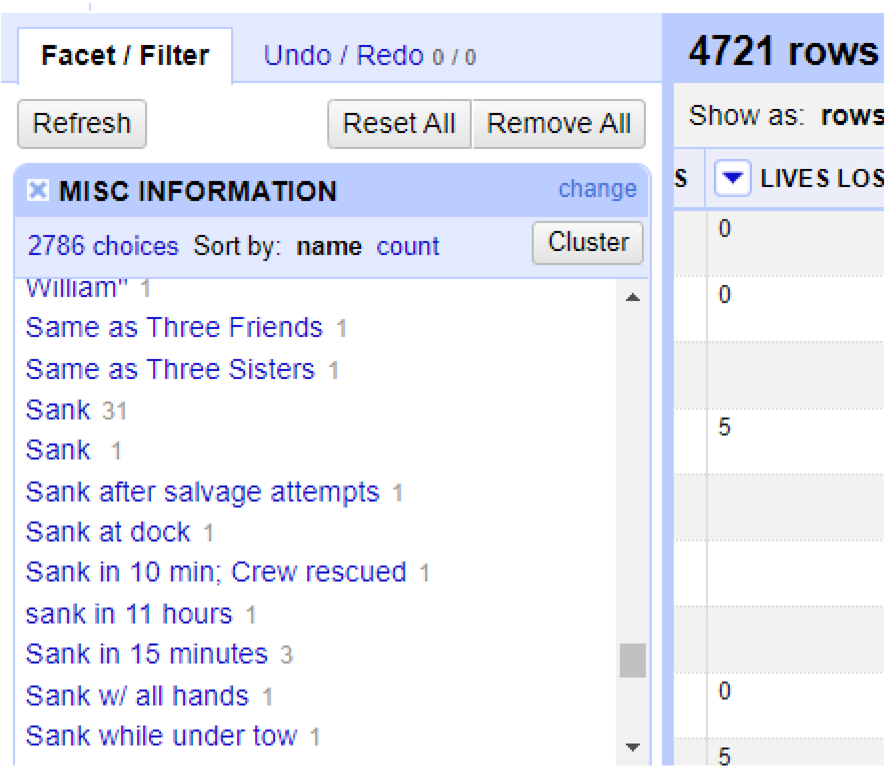
Information about ship measurement and Information caused by the damage are stored in the same table.

The "MISC INFORMATION" column can be made more useful for further analysis following the following steps:

- Created a Text Facet on the "MISC INFORMATION" column, which showed "Sank" count as 31 and "Sank " count as 1.



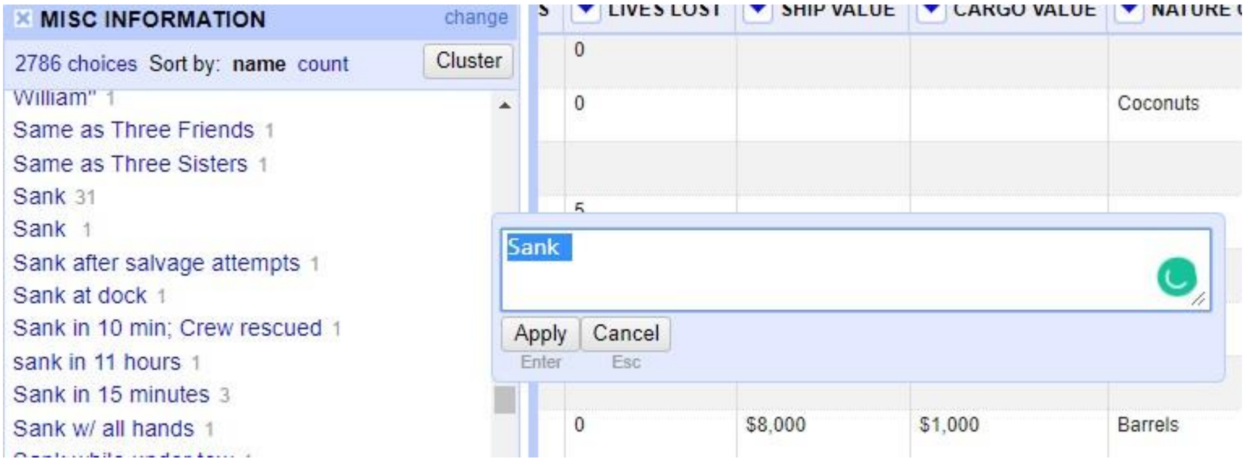
Sank 31
Sank 1



2786 choices Sort by: name count

William" 1
Same as Three Friends 1
Same as Three Sisters 1
Sank 31
Sank 1
Sank after salvage attempts 1
Sank at dock 1
Sank in 10 min; Crew rescued 1
sank in 11 hours 1
Sank in 15 minutes 3
Sank w/ all hands 1
Sank while under tow 1

LIVES LOS
0
0
5
0
5



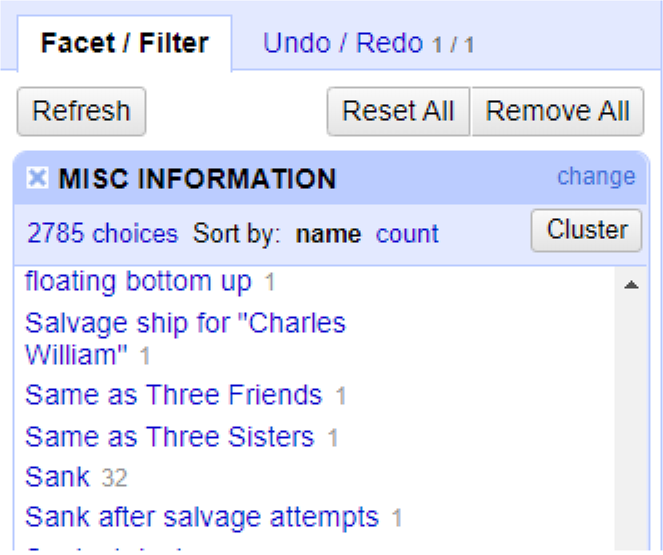
2786 choices Sort by: name count

William" 1
Same as Three Friends 1
Same as Three Sisters 1
Sank 31
Sank 1
Sank after salvage attempts 1
Sank at dock 1
Sank in 10 min; Crew rescued 1
sank in 11 hours 1
Sank in 15 minutes 3
Sank w/ all hands 1

LIVES LOST	SHIP VALUE	CARGO VALUE	NATURE
0			
0			Coconuts
5			
0	$8,000	$1,000	Barrels

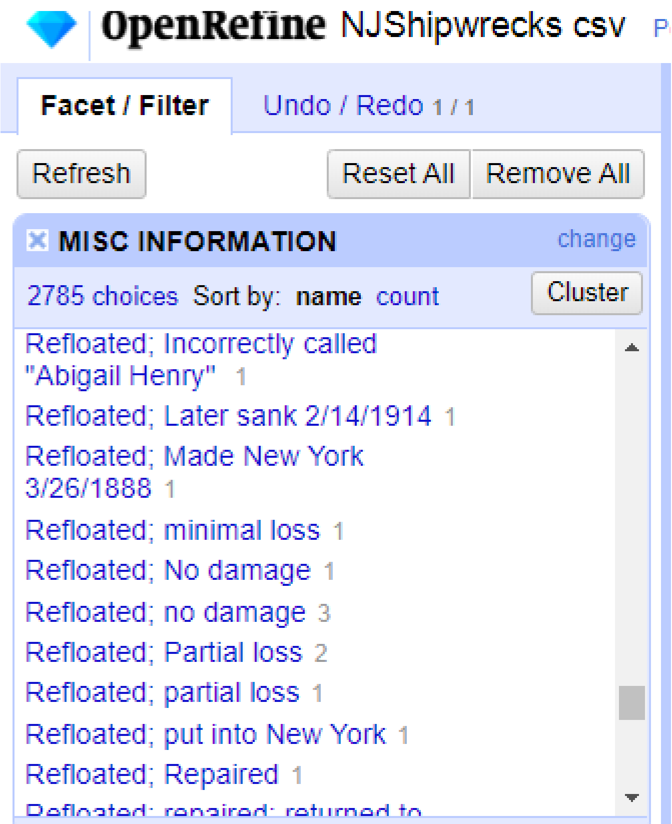
Sank

After removing space from Sank it got merged with Sank and then total there were 32 Sank entries:



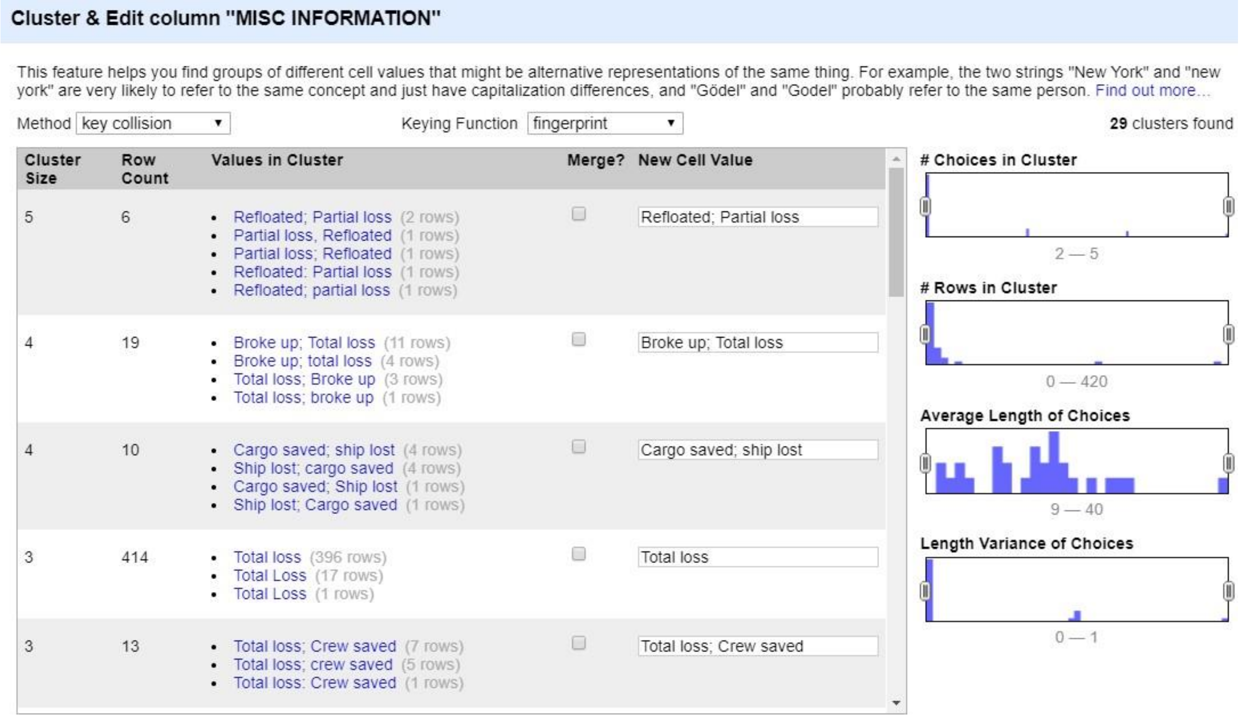
floating bottom up 1
Salvage ship for "Charles William" 1
Same as Three Friends 1
Same as Three Sisters 1
Sank 32
Sank after salvage attempts 1

- If we click on the cluster as shown below:



Refloated; Incorrectly called "Abigail Henry" 1
Refloated; Later sank 2/14/1914 1
Refloated; Made New York 3/26/1888 1
Refloated; minimal loss 1
Refloated; No damage 1
Refloated; no damage 3
Refloated; Partial loss 2
Refloated; partial loss 1
Refloated; put into New York 1
Refloated; Repaired 1
Refloated; repaired; returned to

So, there were 29 clusters found:



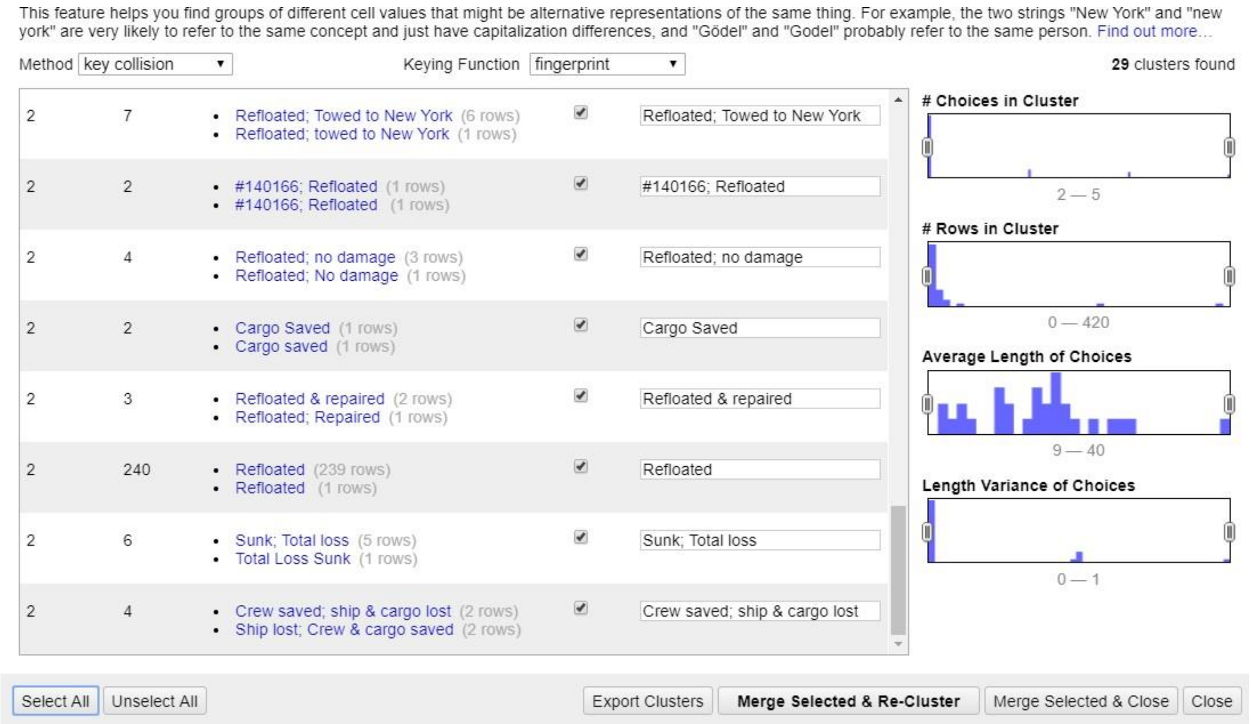
Cluster & Edit column "MISC INFORMATION"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more…

Method: key collision — Keying Function: fingerprint — **29** clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
5	6	Refloated; Partial loss (2 rows) Partial loss, Refloated (1 rows) Partial loss; Refloated (1 rows) Refloated: Partial loss (1 rows) Refloated; partial loss (1 rows)	☐	Refloated; Partial loss
4	19	Broke up; Total loss (11 rows) Broke up; total loss (4 rows) Total loss; Broke up (3 rows) Total loss; broke up (1 rows)	☐	Broke up; Total loss
4	10	Cargo saved; ship lost (4 rows) Ship lost; cargo saved (4 rows) Cargo saved; Ship lost (1 rows) Ship lost; Cargo saved (1 rows)	☐	Cargo saved; ship lost
3	414	Total loss (396 rows) Total Loss (17 rows) Total Loss (1 rows)	☐	Total loss
3	13	Total loss; Crew saved (7 rows) Total loss; crew saved (5 rows) Total loss: Crew saved (1 rows)	☐	Total loss; Crew saved

Clicked on Select All



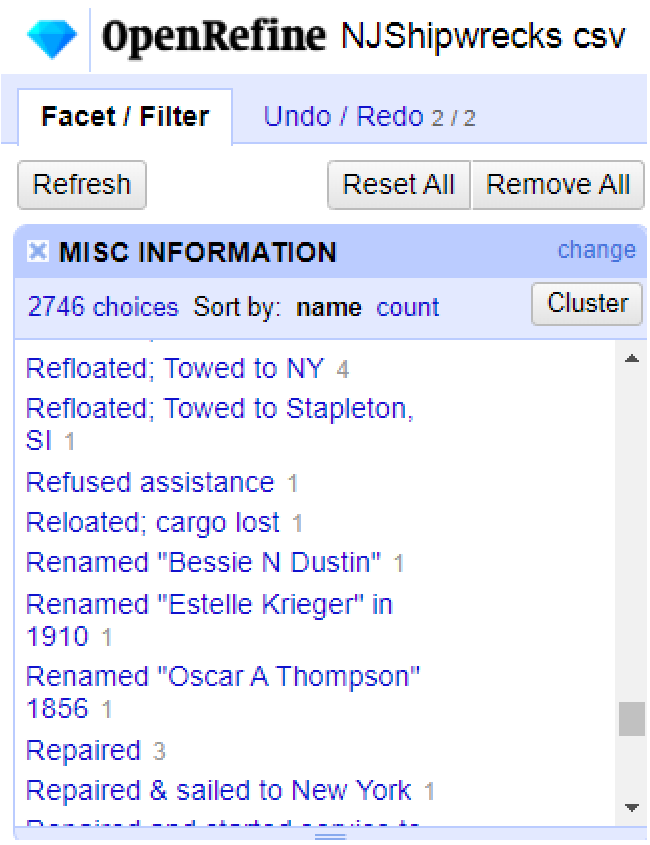
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more…

Method: key collision — Keying Function: fingerprint — **29** clusters found

2	7	Refloated; Towed to New York (6 rows) Refloated; towed to New York (1 rows)	☑	Refloated; Towed to New York
2	2	#140166; Refloated (1 rows) #140166; Refloated (1 rows)	☑	#140166; Refloated
2	4	Refloated; no damage (3 rows) Refloated; No damage (1 rows)	☑	Refloated; no damage
2	2	Cargo Saved (1 rows) Cargo saved (1 rows)	☑	Cargo Saved
2	3	Refloated & repaired (2 rows) Refloated; Repaired (1 rows)	☑	Refloated & repaired
2	240	Refloated (239 rows) Refloated (1 rows)	☑	Refloated
2	6	Sunk; Total loss (5 rows) Total Loss Sunk (1 rows)	☑	Sunk; Total loss
2	4	Crew saved; ship & cargo lost (2 rows) Ship lost; Crew & cargo saved (2 rows)	☑	Crew saved; ship & cargo lost

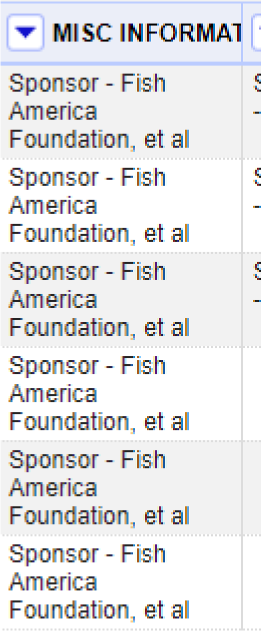
Clicked on Merge Selected & Re-cluster and then Close.

It resulted in the reduction of choice to 2746 from 2785 as shown below and merged semantically similar rows, variant cases rows together:



Refloated; Towed to NY 4
Refloated; Towed to Stapleton, SI 1
Refused assistance 1
Reloated; cargo lost 1
Renamed "Bessie N Dustin" 1
Renamed "Estelle Krieger" in 1910 1
Renamed "Oscar A Thompson" 1856 1
Repaired 3
Repaired & sailed to New York 1

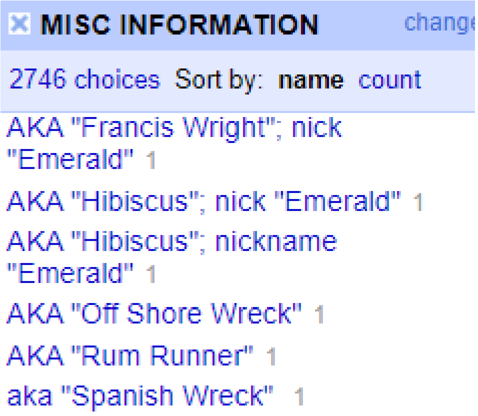
- Many rows have sponsor information as shown below:



MISC INFORMAT
Sponsor - Fish America Foundation, et al
Sponsor - Fish America Foundation, et al
Sponsor - Fish America Foundation, et al
Sponsor - Fish America Foundation, et al
Sponsor - Fish America Foundation, et al
Sponsor - Fish America Foundation, et al

It could be entered in a new column with the name 'Sponsor'.

- There is also aka information mentioned which could be written in AKA column for the rows below:



AKA "Francis Wright"; nick "Emerald" 1

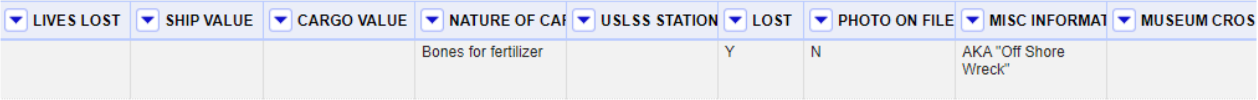
AKA "Hibiscus"; nick "Emerald" 1

AKA "Hibiscus"; nickname "Emerald" 1

AKA "Off Shore Wreck" 1

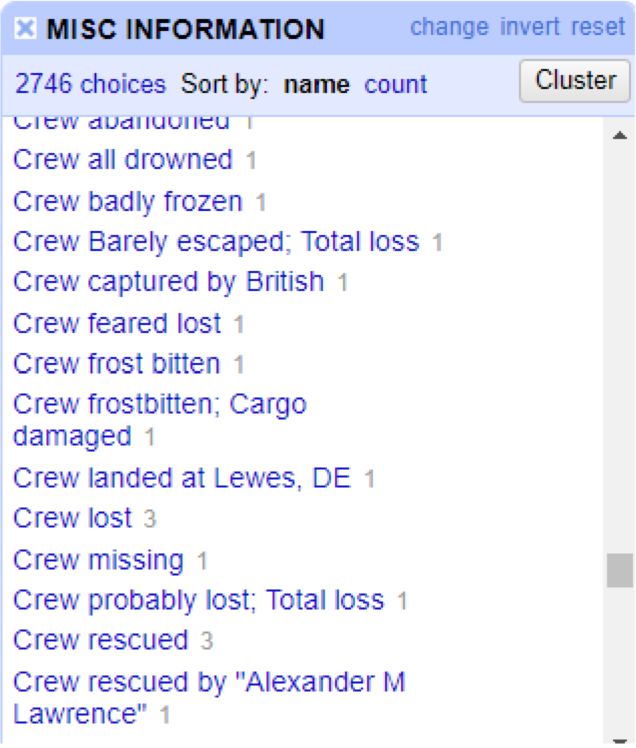
AKA "Rum Runner" 1

aka "Spanish Wreck" 1



LIVES LOST	SHIP VALUE	CARGO VALUE	NATURE OF CAI	USLSS STATION	LOST	PHOTO ON FILE	MISC INFORMAT	MUSEUM CROS
			Bones for fertilizer		Y	N	AKA "Off Shore Wreck"	

- Crew Information can be added in a separate column, which would give information about the condition of the crew.



Crew abandoned 1

Crew all drowned 1

Crew badly frozen 1

Crew Barely escaped; Total loss 1

Crew captured by British 1

Crew feared lost 1

Crew frost bitten 1

Crew frostbitten; Cargo damaged 1

Crew landed at Lewes, DE 1

Crew lost 3

Crew missing 1

Crew probably lost; Total loss 1

Crew rescued 3

Crew rescued by "Alexander M Lawrence" 1