

DATA 201 – Assignment 3

Alex Stevenson – 30073617

Cleaning 1: Unnecessary Punctuation Marks

Before:

event	venue	place	physical_description	occasion	notes
BREAKFAST	COMMERCIAL	HOT SPRINGS, AR	CARD; 4.75X7.5;	EASTER;	
[DINNER]	COMMERCIAL	MILWAUKEE, [WI];	CARD; ILLUS; COL; 7.0X9.0;	EASTER;	WEDGEWOOD BLUE CARD;
FRUHSTUCK/BREAKFAST;	COMMERCIAL	DAMPFER KAISER WILHELM DER C	CARD; ILLU; COL; 5.5X8.0;		MENU IN GERMAN AND EN
LUNCH;	COMMERCIAL	DAMPFER KAISER WILHELM DER C	CARD; ILLU; COL; 5.5X8.0;		MENU IN GERMAN AND EN
DINNER;	COMMERCIAL	DAMPFER KAISER WILHELM DER C	FOLDER; ILLU; COL; 5.5X7.5;		MENU IN GERMAN AND EN
PANY [DINNER]	COMMERCIAL	R.M.S. EMPRESS OF CHINA	CARD; ILLUS; COL; 4.75X7.25;		ILLUS, RED AND WHITE CHE
SUPPER	COMMERCIAL	NEW YORK, [NY];	CARD; ILLUS; COL; 6.0X8.75;		HOTEL CREST IN BLUE; PRIC
FRUHSTUCK/BREAKFAST	COMMERCIAL	SCHNEL DAMPFER KAISER WILHE	BROADSIDE; ILLUS; COL; 5.5X8.50;		MENU IN GERMAN AND EN
LUNCH	COMMERCIAL	DAMPFER KAISER WILHELM DER C	BROADSIDE; ILLUS; COL; 5.5X8.50;		MENU IN GERMAN AND EN
[DINNER]	COMMERCIAL	DAMPFER KAISER WILHELM DER C	FOLDER; ILLUS; COL; 5.5X7.5;		MENU IN GERMAN AND EN
CAFE LUNCHEON	COMMERCIAL	[NEW YORK, NY]	CARD; ILLUS; COL; 4.25X5.5;		HOTEL CREST IN BLUE;

After:

event	venue	place	physical_description	occasion	notes
BREAKFAST	COMMERCIAL	HOT SPRINGS, AR	CARD; 4.75X7.5;	EASTER	
DINNER	COMMERCIAL	MILWAUKEE, WI	CARD; ILLUS; COL; 7.0X9.0;	EASTER	WEDGEWOOD BLUE CARD;
FRUHSTUCK/BREAKFAST	COMMERCIAL	DAMPFER KAISER WILHELM DER C	CARD; ILLU; COL; 5.5X8.0;		MENU IN GERMAN AND EN
LUNCH	COMMERCIAL	DAMPFER KAISER WILHELM DER C	CARD; ILLU; COL; 5.5X8.0;		MENU IN GERMAN AND EN
DINNER	COMMERCIAL	DAMPFER KAISER WILHELM DER C	FOLDER; ILLU; COL; 5.5X7.5;		MENU IN GERMAN AND EN
PANY DINNER	COMMERCIAL	R.M.S. EMPRESS OF CHINA	CARD; ILLUS; COL; 4.75X7.25;		ILLUS, RED AND WHITE CHE
SUPPER	COMMERCIAL	NEW YORK, NY	CARD; ILLUS; COL; 6.0X8.75;		HOTEL CREST IN BLUE; PRIC
FRUHSTUCK/BREAKFAST	COMMERCIAL	SCHNEL DAMPFER KAISER WILHE	BROADSIDE; ILLUS; COL; 5.5X8.50;		MENU IN GERMAN AND EN
LUNCH	COMMERCIAL	DAMPFER KAISER WILHELM DER C	BROADSIDE; ILLUS; COL; 5.5X8.50;		MENU IN GERMAN AND EN
DINNER	COMMERCIAL	DAMPFER KAISER WILHELM DER C	FOLDER; ILLUS; COL; 5.5X7.5;		MENU IN GERMAN AND EN
CAFE LUNCHEON	COMMERCIAL	NEW YORK, NY	CARD; ILLUS; COL; 4.25X5.5;		HOTEL CREST IN BLUE;

Data Quality Issue: Data Integration Error

It looks like some sources of data use semicolons to mark the end of lines, and square brackets to highlight certain elements. There are many semicolons at the end of certain points of data that seem to have come from the same source, while various data points are surrounded by square brackets that could denote a differing source of data. An example of this would be the difference between “[DINNER], DINNER; , and LUNCH”

Why these entities need to be cleaned:

These entities should be cleaned as their presence makes it more difficult to read and compare data. As with the above example, different tuples contain the same information (“DINNER” and “[DINNER]”) should be formatted the same way. Cleaning these entities will make them clearer without erroneous punctuation marks, make them easier to group together as identical elements will be formatted the same way, and will be easier to analyze and compare different rows that have the same elements.

Steps to Clean: (Excel)

1. Select each column not including **physical_description** and **notes**.
2. Ctrl-H to open the Replace menu, find “[” and replace it with nothing

B	C	D	E	F	G	H	I
te	sponsor	event	venue	place	physical_description	occasion	notes
	HOTEL EASTMAN	BREAKFAST	COMMERCIAL	HOT SPRINGS, AR	CARD; 4.75X7.5;	EASTER;	
	REPUBLICAN HOUSE	[DINNER]	COMMERCIAL	MILWAUKEE, [WI];	CARD; ILLU; COL; 7.0X9.0;	EASTER;	WEDGEWOOD BLUE CARL
	NORDDEUTSCHER LLOYD BREMEN	FRUHSTUCK/BREAKFAST;	COMMERCIAL	DAMPFER KAISER WILHELM DER C	CARD; ILLU; COL; 5.5X8.0;		MENU IN GERMAN AND E
	NORDDEUTSCHER LLOYD BREMEN	LUNCH;	COMMERCIAL	DAMPFER KAISER WILHELM DER C	CARD; ILLU; COL; 5.5X8.0;		MENU IN GERMAN AND E
	NORDDEUTSCHER LLOYD BREMEN	DINNER;			5X7.5;		MENU IN GERMAN AND E
	CANADIAN PACIFIC RAILWAY COMPANY	[DINNER]			5X7.25;		ILLUS, RED AND WHITE C
	HOTEL NETHERLAND	SUPPER			8.75;		HOTEL CREST IN BLUE; PR
	NORDDEUTSCHER LLOYD BREMEN	FRUHSTUCK/BREAKFAST			IL; 5.5X8.50;		MENU IN GERMAN AND E
	NORDDEUTSCHER LLOYD BREMEN	LUNCH			IL; 5.5X8.50;		MENU IN GERMAN AND E
	NORDDEUTSCHER LLOYD BREMEN	[DINNER]			.5X7.5;		MENU IN GERMAN AND E
	HOTEL MARLBOROUGH	CAFE LUNCHEON			5X5.5;		HOTEL CREST IN BLUE;
	ALPHA OF ZETA PSI	ANNUAL BANQUET			5.5X7.0;		VELLUM COVER; CREST O
	MANHATTAN HOTEL	DINNER					A LA CARTE DU JOUR; HO
	PACIFIC MAIL STEAMSHIP COMPANY	DINNE					DECORATIVE BORDER;
	OCCIDENTAL & ORIENTAL	BREAKFAST	COMMERCIAL	S.S. "DORIC"	BROADSIDE; ILLU; 5.5X8.5;		HANDWRITTEN; STEAMSH

3. Do the same with “]” and “;”

Cleaning 2: Empty Columns

Before:

es	call_number	keywords	language	date	location	location_type	currency	currency_symbol	status	page_count	dis
	1900-2822			4/15/1900	Hotel Eastman				complete	2	67
WEDGEWOOD BLUE CARD, WHITE EMBOSSED GREEK KEY BORDER, "EASTER SUNDAY" EMBOSSED IN WHITE. VIOLET COLORED SPRAY OF FLOWERS IN UPPER LEFT CORNER;	1900-2825			4/15/1900	Republican House				complete	2	34
MENU IN GERMAN AND ENGLISH; ILLUS. STEAMSHIP AND SAILING VESSEL.	1900-2827			4/16/1900	Norddeutscher Lloyd Bremen				complete	2	84
MENU IN GERMAN AND ENGLISH; ILLUS. HARBOR SCENE WITH SAILING VESSEL.	1900-2828			4/16/1900	Norddeutscher Lloyd Bremen				complete	2	63
MENU IN GERMAN AND ENGLISH; ILLUS. HARBOR SCENE WITH ROCKS AND LIGHTHOUSE. STEAMSHIP AND SAILING VESSELS; CONCERT PROGRAM DATES. ON GERMAN SIDE OF MENU "MONTAG, DEN 16 APRIL 1900"; ON ENGLISH SIDE OF MENU "MONDAY, APRIL 15TH, 1900";	1900-2829			4/16/1900	Norddeutscher Lloyd Bremen				complete	4	33
ILLUS. RED AND WHITE CHECKERED FLAG.	1900-2831			4/16/1900	Canadian Pacific Railway Company				complete	2	37
HOTEL CREST IN BLUE; PRICED MENU.	1900-2838			4/16/1900	Hotel Netherland		Dollars	\$	complete	2	144
MENU IN GERMAN AND ENGLISH; ILLUS. LIGHTHOUSE. STEAMSHIP	1900-2839			4/17/1900	Norddeutscher Lloyd Bremen				complete	2	80

After:

Facet / Filter	Undo / Redo 4 / 4	2000 rows									Extensions		Wikidata		
Extract... Apply...		Show as: rows records	Show: 5 10 25 50 100 500 1000 rows							« first		« previous	1 2	next »	last »
Filter		physical_description	occasion	notes	call_number	date	location	currency	currency_symbol	status	dis				
0. Create project		ILLUS. COL: 4.75X7.5;	EASTER;		1900-2822	4/15/1900	Hotel Eastman			complete	2				
1. Remove column keywords		ILLUS. COL: 7.0X9.0;	EASTER;	WEDGEWOOD BLUE CARD, WHITE EMBOSSED GREEK KEY BORDER, "EASTER SUNDAY" EMBOSSED IN WHITE, VIOLET COLORED SPRAY OF FLOWERS IN UPPER LEFT CORNER;	1900-2825	4/15/1900	Republican House			complete	2				
2. Remove column language		ILLU. COL: 5.5X8.0;		MENU IN GERMAN AND ENGLISH; ILLUS. STEAMSHIP AND SAILING VESSEL.	1900-2827	4/16/1900	Norddeutscher Lloyd Bremen			complete	2				
3. Remove column location_type		ILLU. COL: 5.5X8.0;		MENU IN GERMAN AND ENGLISH; ILLUS. HARBOR SCENE WITH SAILING VESSEL.	1900-2828	4/16/1900	Norddeutscher Lloyd Bremen			complete	2				
4. Remove column name		ER; ILLU. COL: 5.5X7.5;		MENU IN GERMAN AND ENGLISH; ILLUS. HARBOR SCENE WITH ROCKS AND LIGHTHOUSE. STEAMSHIP AND SAILING VESSELS; CONCERT PROGRAM DATES. ON GERMAN SIDE OF MENU "MONTAG, DEN 16 APRIL 1900"; ON ENGLISH SIDE OF MENU "MONDAY, APRIL 15TH, 1900";	1900-2829	4/16/1900	Norddeutscher Lloyd Bremen			complete	4				
		ILLUS. COL: 4.75X7.25;		ILLUS. RED AND WHITE CHECKERED FLAG.	1900-2831	4/16/1900	Canadian Pacific Railway Company			complete	2				
		ILLUS. COL: 6.0X8.75;		HOTEL CREST IN BLUE; PRICED MENU.	1900-2838	4/16/1900	Hotel Netherland	Dollars	\$	complete	2				
		DSIDE; ILLUS. COL: 5.5X8.50;		MENU IN GERMAN AND ENGLISH; ILLUS. LIGHTHOUSE. STEAMSHIP	1900-2839	4/17/1900	Norddeutscher Lloyd Bremen			complete	2				

Data Quality Issue: Missing Data

It looks like the data set was supposed to contain the attributes **name**, **keywords**, **language**, and **location_type**. At some point this data was lost or removed after being collected as zero entries have any of these fields populated.

Why these entities need to be cleaned:

These columns are unnecessary and clutter up the data with random empty attributes, making it more difficult to read. Particularly as there are so many other fields that could be important to a user. Additionally, if you are taking a look at a small amount of rows at once, the existence of these columns implies that other rows will have that data available and that the current rows are not complete.

Steps to Clean: (OpenRefine)

1. Click the name of an unnecessary column (**name**, **keywords**, **language**, and **location_type**)
2. Click **Edit Column**
3. Click **Remove This Column**

call_number	keywords	language	date	location	location_type
1900-2822	Facet		4/15/1900	Hotel Eastman	
1900-2825	Text filter		4/15/1900	Republican House	
	Edit cells				
1900-2827	Edit column				
	Transpose				
1900-2828	Sort...				
	View				
1900-2829	Reconcile				
1900-2831					
1900-2838					
1900-2839					

Split into several columns...

Join columns...

Add column based on this column...

Add column by fetching URLs...

Add columns from reconciled values...

Rename this column...

Remove this column

Move column to beginning

Move column to end

Move column left

Move column right

Cleaning 3: Multiple Values for Unknown Fields

Before:

MENU	COMMERCIAL		FOLDER; 6X12;		A LA CARTE MEN
MENU	COMMERCIAL	NEW YORK, NY	FOLDER; 6.75X10;		A LA CARTE MEN
[FAREWELL DINNER GIVEN E	OTHER (PRIVATE PAR	SHERRY'S, NEW YORK, NY	CARD; ILLUS; 4.5X7;	COMPLIMENTARY/TE	FRENCH; HOST'S
DE BUSINESS MEN'S LUNCH	COMMERCIAL	NEW YORK, NY	CARD; ILLUS; 4.5X5.75;		TABLE D'HOTE U
DE TABLE D'HOTE DINNER	COMMERCIAL	NEW YORK, NY	CARD; ILLUS; 4.5X5.75;		TABLE D'HOTE D
MENU	COMMERCIAL	NEW YORK, NY	BOOKLET; ILLUS; 6X9.5;		PRICED MENU; I
SPRING DINNER	(SOC?);	ST. DENIS HOTEL	BOOKLET; ILLUS; 5.5X7.75;	OTHER (ANNUAL DINI	INCLUDES PHOT
BREAKFAST	COMMERCIAL	EN ROUTE ABOARD R.M.S. LUCAN	CARD; 4.25 X 6.5;		LOGO; WINE LIS
11TH ANNUAL BANQUET	POL;	MANHATTAN HOTEL, NY	FOLDER; 5 X 6.5;	78TH BIRTHDAY OF U.	LIST OF WINES S
OF 25TH ANNUAL MEETING & EDUC,	?	IROQUOIS	FOLDER; COL; 5 X 7;	ANNUAL	PROGRAM & SPE
A. BANQUET	?	?	FOLDER; ILLUS; 4.5 X 6.5;	?	
BREAKFAST	COMMERCIAL	WASHINGTON, D.C.	BROADSIDE; COL; 4.5 X 7;	DAILY	
DINNER	COMMERCIAL	42ND ST. & MAD. AVE, NY	BROADSIDE; ILLUS; 6 X 9.75	DAILY	
BREAKFAST	COMMERCIAL	EN ROUTE	BROADSIDE; ILLUS; COL; 5 X 8;	DAILY;	MENU HANDWR
LUNCH	COMMERCIAL	EN ROUTE	BROADSIDE; ILLUS; COL; 5 X 8;	DAILY	MENU HANDWR
LUNCH OR DINNER	COMMERCIAL	TAOEMINA, SICILY	BROADSIDE; 3.75 X 6;		MENU IN ITALIA

After:

Find and Replace

Find what: ?

Replace with:

Within: Sheet

Search: By Rows

Look in: Formulas

Match case

Match entire cell contents

Options <<

Replace All

Microsoft Office Excel

Excel has completed its search and has made 2614 replacements.

OK

Data Quality Issue: Entity Resolution

Some unknown entries in the database are labelled with “?”, while others are left blank. You can also see that this issue is not consistent per row, as some rows have several question marks along with several empty attributes. This is an entity resolution error as there are multiple different entries that mean the same thing. I don’t think this is a data integration error as the error is not consistent within individual rows.

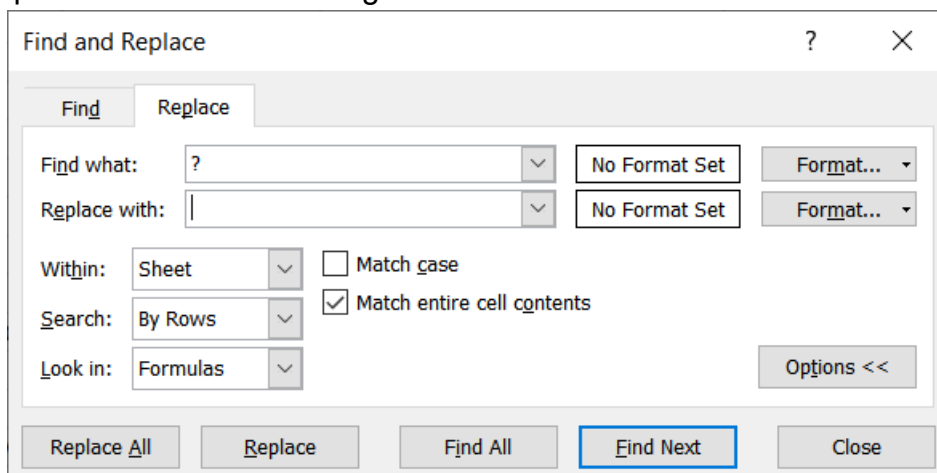
Why these entities need to be cleaned:

These entities consisting of "?" should simply be deleted as their presence adds nothing to the dataset but inconsistency. There are many blank fields that denote information that is not available, two different formats to explain the same thing is unnecessary and confusing. Removal of these entities should help with analysis by reducing confusion, making the format of unknown fields consistent, and make it obvious when a field is unknown without a dubious question mark.

Note that in the screenshots above, the entry (SOC?); is highlighted in yellow. This is to make sure that not all question marks are removed, only the ones that fill the entire cell on their own.

Steps to Clean: (Excel)

1. Ctrl-H to open the Replace menu
2. Find "?" and replace it with nothing
3. Click **Options** to see the advanced options
4. Make sure that **Match entire cell contents** is checked, this will prevent the necessary question marks from being removed



5. Click **Replace All**

Cleaning 4: Breaking Big Attributes into Multiple Columns

Before:

place	physical_description	occasion	notes	call_number
HOT SPRINGS, AR	CARD; 4.75X7.5;	EASTER;		1900-2822
MILWAUKEE, [WI];	CARD; ILLUS; COL; 7.0X9.0;	EASTER;	WEDGEWOOD BLUE CARD; WHITE EMBOSSED GREEK KEY BORDER; "EASTER SUNDAY" EMBOSSED IN WHITE; VIOLET COLORED SPRAY OF FLOWERS IN UPPER LEFT CORNER;	1900-2825
DAMPFER KAISER WILHELM DER GROSSE;	CARD; ILLU; COL; 5.5X8.0;		MENU IN GERMAN AND ENGLISH; ILLUS. STEAMSHIP AND SAILING VESSEL;	1900-2827
DAMPFER KAISER WILHELM DER GROSSE;	CARD; ILLU; COL; 5.5X8.0;		MENU IN GERMAN AND ENGLISH; ILLUS. HARBOR SCENE WITH SAILING VESSEL;	1900-2828
DAMPFER KAISER WILHELM DER GROSSE;	FOLDER; ILLU; COL; 5.5X7.5;		MENU IN GERMAN AND ENGLISH; ILLUS. HARBOR SCENE WITH ROCKS AND LIGHTHOUSE; STEAMSHIP AND SAILING VESSELS; CONCERT PROGRAM; DATES: ON GERMAN SIDE OF MENU "MONTAG, DEN 16 APRIL 1900"; ON ENGLISH SIDE OF MENU "MONDAY, APRIL 15TH, 1900";	1900-2829
R.M.S. EMPRESS OF CHINA	CARD; ILLUS; COL; 4.75X7.25;		ILLUS. RED AND WHITE CHECKERED FLAG;	1900-2831
NEW YORK, [NY];	CARD; ILLUS; COL; 6.0X8.75;		HOTEL CREST IN BLUE; PRICED MENU;	1900-2838
SCHNELLDAMPFER KAISER WILHELM DER GROSSE;	BROADSIDE; ILLUS; COL; 5.5X8.50;		MENU IN GERMAN AND ENGLISH; ILLUS. LIGHTHOUSE; STEAMSHIP; FISHING DORY;	1900-2839
DAMPFER KAISER WILHELM DER GROSSE;	BROADSIDE; ILLUS; COL; 5.5X8.50;		MENU IN GERMAN AND ENGLISH; ILLUS. SAILING SHIP/SHEETS FURLED; STEAMSHIP;	1900-2840
DAMPFER KAISER WILHELM DER GROSSE;	FOLDER; ILLUS; COL; 5.5X7.5;		MENU IN GERMAN AND ENGLISH; ILLUS. HARBOR; LIGHTHOUSE; ROCKS; STEAMSHIP; SAILING SHIPS; CONCERT PROGRAM;	1900-2841
[NEW YORK, NY]	CARD; ILLUS; COL; 4.25X5.5;		HOTEL CREST IN BLUE;	1900-2843
DELMONICO'S, [NEW YORK, NY];	BOOKLET; ILLUS; COL; 5.5X7.0;		VELLUM COVER; CREST OF ZETA PSI; TIED WITH BLUE SATIN RIBBON; PRICED WINE LIST; ALPHA CHAPTER OFFICERS; BANQUET COMMITTEE; TOASTS; LYRICS TO "ZETA PSI WE	1900-2844

After:

scription 3	physical_description 4	physical_description 5	occasion	notes 1	notes 2	notes 3	notes 4	notes 5	notes 6	notes 7
	7.0X9.0;		EASTER;	WEDGEWOOD BLUE CARD	WHITE EMBOSSED GREEK KEY BORDER	"EASTER SUNDAY" EMBOSSED IN WHITE	VIOLET COLORED SPRAY OF FLOWERS IN UPPER LEFT CORNER;			
	5.5X8.0;			MENU IN GERMAN AND ENGLISH	ILLUS. STEAMSHIP AND SAILING VESSEL;					
	5.5X8.0;			MENU IN GERMAN AND ENGLISH	ILLUS. HARBOR SCENE WITH SAILING VESSEL;					
	5.5X7.5;			MENU IN GERMAN AND ENGLISH	ILLUS. HARBOR SCENE WITH ROCKS AND LIGHTHOUSE	STEAMSHIP AND SAILING VESSELS	CONCERT PROGRAM	DATES: ON GERMAN SIDE OF MENU "MONTAG, DEN 16 APRIL 1900"	ON ENGLISH SIDE OF MENU "MONDAY, APRIL 15TH, 1900";	
	4.75X7.25;			ILLUS. RED AND WHITE CHECKERED FLAG;						
	6.0X8.75;			HOTEL CREST IN BLUE	PRICED MENU;					
	5.5X8.50;			MENU IN GERMAN AND ENGLISH	ILLUS. LIGHTHOUSE	STEAMSHIP	FISHING DORY;			
	5.5X8.50;			MENU IN GERMAN AND ENGLISH	ILLUS. SAILING SHIP/SHEETS FURLED	STEAMSHIP;				
	5.5X7.5;			MENU IN GERMAN AND ENGLISH	ILLUS. HARBOR	LIGHTHOUSE	ROCKS	STEAMSHIP	SAILING SHIPS	CONCERT PROGRAM;
	4.25X5.5;			HOTEL CREST IN BLUE;						
	5.5X7.0;			VELLUM COVER	CREST OF ZETA PSI	TIED WITH BLUE SATIN RIBBON	PRICED WINE LIST	ALPHA CHAPTER OFFICERS	BANQUET COMMITTEE	TOASTS

Data Quality Issue: Entity Resolution

This is a entity resolution issue because the data has been collected, but is stored in a confusing and difficult to parse way. Both **physical_description** and **notes** have several entries for a majority of different rows, however all of those entries are listed off in a impenetrable block of capitalized text.

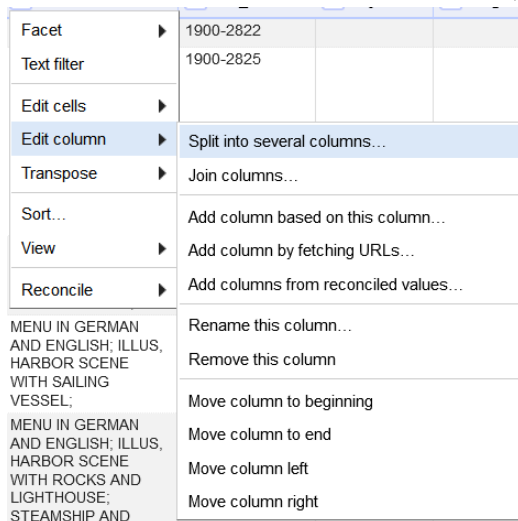
Why these entities need to be cleaned:

It should be useful to break these two fields into multiple columns. Both of them consistently have 2+ entries separated by semicolons, which makes it difficult to read and understand the data provided. By separating each entry into its own column, it will become clearer and easier to compare the different tuples.

As it is now, users have to read the entire line of notes/physical descriptions in order to find specific ones that they are looking for as it is just a line of capitalized text. By breaking it into their own columns you are able to skim those fields and quickly find what you are looking for, while also being able to compare different rows more easily.

Steps to Clean: (OpenRefine)

1. Click the name of the column to edit, **Edit Column**, and **Split into several columns**



2. The separator should be “;” as that is what denotes the separation of list elements in the data

A screenshot of the 'Split column notes into several columns' dialog box in OpenRefine. The dialog has two main sections: 'How to split column' and 'After Splitting'. In the 'How to split column' section, the 'by separator' radio button is selected. The 'Separator' field contains a semicolon (;). The 'Split into' field is set to 'columns at most (leave blank for no limit)'. In the 'After Splitting' section, the 'Guess cell type' and 'Remove this column' checkboxes are both checked. At the bottom, there are 'OK' and 'Cancel' buttons. A small text box at the bottom left says 'List of integers separated by commas, e.g., 5, 7, 15'.

Cleaning 5: Differing Date Formats

Before:

	2/19/1900	Marie Antoinette Hotel
	2/20/1900	Red Star Line S.S.Southwark
	2/20/1900	Marie Antoinette Hotel
	2/20/1900	Third Panel Sheriff's Jury New York County
	1888-10-15	The Albany
	1865-09-28	Parker House

After:

	1900-02-19	Marie Antoinette Hotel
	1900-02-20	Red Star Line S.S.Southwark
	1900-02-20	Marie Antoinette Hotel
	1900-02-20	Third Panel Sheriff's Jury New York County
	1888-10-15	The Albany
	1865-09-28	Parker House

Data Quality Issue: Data Integration

It looks like there are two different date formats used in this data set. This implies either that the data has come from at least two different sources that use different date formats, or that the schema has changed over time to use a different date format.

Why these entities need to be cleaned:

The current date formats can be confusing to readers when they look over multiple tuples with different formats. In the screenshots above we see two different date formats in use. This requires more work to parse the data as a human reader or a program, as they will need to account for both formats.

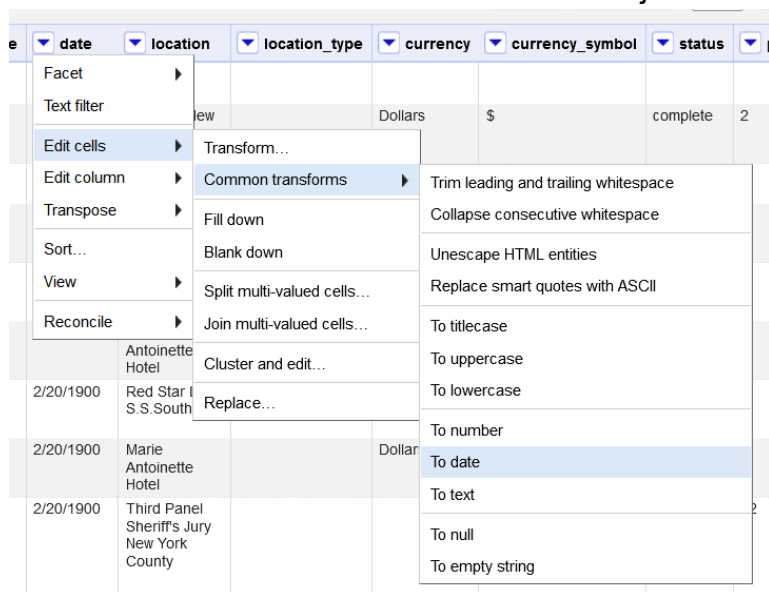
There are also potential uncertainties between the ordering of months and days in the data. An example of this is 9/5/2000, that could denote either May 9th or September 5th.

These different date formats should all be made consistent to prevent confusion and uncertainty as to what the data means.

Additionally, leading zeros should be added to the month and day sections of the dates, to make each date uniform and easier to compare with one another as the length of the dates will all be consistent

Steps to Clean: (OpenRefine)

1. Click the **date** column header, select **Edit Cells**, **Common Transforms**, **To date**, to transform the data in the column to a date object.



2. Do the same as step 1, but transform the column into text.
This step is necessary as the Dates object seems to store additional text than what is displayed
3. Each date should now look like this:

1900-04-16T00:00:00Z	Cal Pac Col
1900-04-16T00:00:00Z	Hot Net
1900-04-17T00:00:00Z	Not Llo

4. Click the **date** column header, select **Edit Cells** and **Replace**, replacing the string **"T00:00:00Z"** with nothing.

Replace

Find

☐ case insensitive ☐ whole word ☐ regular expression
Leave blank to add the replacement string after each character.
Check "regular expression" to find special characters (new lines, tabulations...) or complex patterns.

Replace with

☐ use \n for new lines, \t for tabulation, \\n for \n, \\t for \t.
If "regular expression" option is checked and finding pattern contains groups delimited with parentheses, \$0 will return the complete string matching the pattern, and \$1, \$2... the 1st, 2nd... group.