Rekisstory Tutorial for Statistical Data Integration by CSV Import

2024-05-14
Go Sugimoto

In this document you will learn how to use the Rekisstory data integration function by CSV import, using a sample statistical data

1 Compare external statistical data with Rekisstory

It is easy to integrate Rekisstory's search results with your data with statistics to perform an interesting scientific analysis.

As an example, we use a publicly available dataset from HISTALP (Historical Instrumental Climatological Surface Time Series Of The Greater Alpine Region). The dataset we will use contains the weather data in Vienna from 1775 to 2024. Unfortunately, the data is not in the same format as Rekisstory, thus, we created a sample data normalised for you. In this case, we updated the temperature values as floats and reorganise the data structure: HISTALP_AT_WIE_T01_1760_2025_cleaned.csv [1] (or download it from the Rekisstory tutorial page)

In future, you probably need to process your data at hand before data import. You may not need to "map" your data, because we have a simple data mapping function. However, it is important to follow our data type format in your data, so we can render your data correctly. This is especially true for geo-coordinates and dates that require specific syntax.

---

[1] "Data source: GeoSphere Austria - https://data.hub.geosphere.at ". See the terms of use from the data source website.

Walk-through

Go to Data Integration page from the top menu bar

Import CSV file (Step 1A)

External Data Integration with Rekisstory

- Here you can integrate external data with our Comparison Search results
- Three methods are available to import data (CSV upload, REST API, or SPARQL)
- Follow Step 1 to Step 6
- Various types of data including statistics should work, but it is best for visualization to have time
- This service is tested with simple structured data and endpoints, complicated settings may not w
- You can start the data integration (Step 1) at any point of this process, but a start-over will rem
  CSV and upload it again
- 🚫 Mandatory field

Step 1A: CSV file upload from your local machine

- You can only upload one CSV file at a time
- Only first 100 rows will be imported

Select a CSV file from your local machine

ファイルを選択　HISTALP_…leaned.csv 🚫　Upload

- Click the Select File button and locate the downloaded CSV file on your local machine.
- Click Upload
- If you see the success message, your file is successfully imported.

Reformat CSV (Step 2A)

Step 2A: Check the imported CSV and extract records

Your can specify two parameters to extract the right part of CSV. This extraction is not perfect. You might find it easier to reformat your CSV file on your local machine first

Check the structure of CSV in Table and Text

For checking purpose, the index column is added on the left, which will be removed afterwards. The index is for hint only.

Type a keyword...

Index	year;jan;feb;mar;apr;may;jun;jul;aug;sep;oct;nov;dec	null
0	1775;-3	6 ;24 ;58 ;72 ;117 ;200 ;191 ;204 ;164 ;104 ;37 ;-18
1	1776;-7	1 ;15 ;54 ;94 ;123 ;171 ;192 ;182 ;129 ;67 ;23 ;-31
2	1777;-4	3 ;-12 ;48 ;68 ;148 ;177 ;181 ;196 ;142 ;75 ;37 ;-16
3	1778;0	2 ;-11 ;48 ;119 ;157 ;171 ;216 ;202 ;129 ;86 ;44 ;26
4	1779;-8	3 ;36 ;59 ;135 ;167 ;164 ;180 ;185 ;158 ;103 ;38 ;35
5	1780;-4	8 ;-33 ;72 ;74 ;149 ;169 ;192 ;190 ;145 ;102 ;36 ;-25
6	1781;-2	2 ;05 ;53 ;102 ;152 ;189 ;202 ;212 ;167 ;77 ;49 ;03
7	1782;-0	2 ;-29 ;47 ;81 ;148 ;193 ;207 ;197 ;139 ;83 ;17 ;07
8	1783;2	0 ;34 ;42 ;104 ;164 ;193 ;209 ;197 ;159 ;93 ;42 ;-34
9	1784;-6	6 ;-18 ;34 ;74 ;165 ;181 ;196 ;193 ;172 ;61 ;40 ;-04

Showing **1** to **10** of **99** results

Previous　1　2　3　…　10　Next

year;jan;feb;mar;apr;may;jun;jul;aug;sep;oct;nov;dec

1775;-3,6 ;2,4 ;5,8 ;7,2 ;11,7 ;20,0 ;19,1 ;20,4 ;16,4 ;10,4 ;3,7 ;-1,8

1776;-7,1 ;1,5 ;5,4 ;9,4 ;12,3 ;17,1 ;19,2 ;18,2 ;12,9 ;6,7 ;2,3 ;-3,1

1777;-4,3 ;-1,2 ;4,8 ;6,8 ;14,8 ;17,7 ;18,1 ;19,6 ;14,2 ;7,5 ;3,7 ;-1,6

- Scroll down to Table view of the CSV file
- Check if there is no error when imported

- Table looks awkward, because the file is not the standard comma separate format. We will correct it

```
1778;0,2 ;-1,1 ;4,8 ;11,9 ;15,7 ;17,1 ;21,6 ;20,2 ;12,9 ;8,6 ;4,4 ;2,6

1779;-8,3 ;3,6 ;5,9 ;13,5 ;16,7 ;16,4 ;18,0 ;18,5 ;15,8 ;10,3 ;3,8 ;3,5

1780;-4,8 ;-3,3 ;7,2 ;7,4 ;14,9 ;16,9 ;19,2 ;19,0 ;14,5 ;10,2 ;3,6 ;-2,5

1781;-2,2 ;0,5 ;5,3 ;10,2 ;15,2 ;18,9 ;20,2 ;21,2 ;16,7 ;7,7 ;4,9 ;0,3
```

Record separator/delimiter

Please specify the separator/delimiter of your records (e.g. "," (comma), ";" (semi-colon))

[;] ⓘ

Skip first rows

Specify the number of rows/lines to skip from the top of the file (to exclude unnecessary rows). The first row after the skip will be used as header. If none, type "0", meaning that the first row will be interpreted as header

[0] ⓘ

[Reformat CSV]

- Change the Record Separator/Delimiter from default comma (",") to semi-colon (";")
- As the first row of the table is header, leave Skip the first low as "0"
- Click Reformat CSV button

Check your records & Mapping Table (Step 3 & 4)

Step 3: Check your CSV, API, or SPARQL records

- If you see an unexpected result, please try again from Step 1

[Type a keyword...]

year	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec
1775	-3,6	2,4	5,8	7,2	11,7	20,0	19,1	20,4	16,4	10,4	3,7	-1,8
1776	-7,1	1,5	5,4	9,4	12,3	17,1	19,2	18,2	12,9	6,7	2,3	-3,1
1777	-4,3	-1,2	4,8	6,8	14,8	17,7	18,1	19,6	14,2	7,5	3,7	-1,6
1778	0,2	-1,1	4,8	11,9	15,7	17,1	21,6	20,2	12,9	8,6	4,4	2,6
1779	-8,3	3,6	5,9	13,5	16,7	16,4	18,0	18,5	15,8	10,3	3,8	3,5
1780	-4,8	-3,3	7,2	7,4	14,9	16,9	19,2	19,0	14,5	10,2	3,6	-2,5
1781	-2,2	0,5	5,3	10,2	15,2	18,9	20,2	21,2	16,7	7,7	4,9	0,3
1782	-0,2	-2,9	4,7	8,1	14,8	19,3	20,7	19,7	13,9	8,3	1,7	0,7
1783	2,0	3,4	4,2	10,4	16,4	19,3	20,9	19,7	15,9	9,3	4,2	-3,4
1784	-6,6	-1,8	3,4	7,4	16,5	18,1	19,6	19,3	17,2	6,1	4,0	-0,4

Showing **1** to **10** of **100** results

Previous | 1 | 2 | 3 | ... | 10 | Next

- Scroll down to the table view
- Would your table look organised now? Do you understand what the statistics are about?
- It is the mean temperature of Vienna from 1775 to 2024. In Rekisstory, we only allow users to import first 100 records, thus, you would not see the records of 2023. On the first row, you see highlighted text as headers, mainly the months of a year. In the first column, you see a series of years. In this tutorial, we would like to map the years and the temperature of January.

- You need to map the correct headers to our target headers in the Mapping Table further below.

Step 4: Mapping Table

- You can specify the mapping between your data (CSV file, API, or SPARQL) and the data table used in our service (predefined)
- Please put "Your Column Header" (i.e. column names used in your CSV, keys in the objects used in JSON API, or SPARQL variables) which would correspond to the column headers in our service ("Target Column Header")
- It is important to use the **exact lables** of your column headers/JSON keys/SPARQL variables, which you can check in the generated preview table above. Watch out spaces, dots, underscores, slashes, capital and lower letters
- **At least one mapping should be defined**. There is no mandatory fields to map to our headers. "final_entity_type_uri" is required to use different colours and/or hyperlinks for entities (the red row). Names/labels are required to display something in the result table. Date (at least "starttime") and coordinate ("coordinate_loc") are required for timeline and map respectively. Statistic is required for mixing (see below)(the purplse rows). The headers missing in the mapping will be simply absent in the result. For better results, map as much as possible
- Data type should be followed for proper rendering of your input data, but it is not strictly controlled during the mapping
- You can use the same header more than once in the mapping
- It is possible to mix our data and statistics in a simplified timeline chart by specifying the statistic column below. A downside is only points in time (no duration data) and label info are shown. If you would like to see duration data, use another column for mapping (Object label etc), but the mixed timeline cannot be generated. Customise the design of the chart further below

Your Column Header	Target Column Header	Data Type	Description	Example
column header 1	final_entity_type_uri	xsd:anyURI (Desirable)	Unique Identifier. URI of entity (Desirable)	123 (integer), "123" (string), http://dbpedia.org/resource/Shohei_Ohtani (Desirable)
year	final_entity_type_name	RDF Literal	Name of entity	Shohei Ohtani
column header 3	object	xsd:anyURI	URI of object	http://www.wikidata.org/entity/statement/Q34661-100ba6c8-43a9-5d4f-62d9-f3b21f08311a
jan	objectLabel	RDF Literal	Name of object	Judith and the Head of Holofernes
column header 5	roleType_uri	xsd:anyURI	URI of predicate	http://www.wikidata.org/entity/P19
column header 6	roleTypeLabel	RDF Literal	Name of predicate	Place of birth
column header 7	image	xsd:anyURI	URI of image for the object	http://commons.wikimedia.org/wiki/Special:FilePath/Judith%201%20%28cropped%29.jpg
year	starttime	xsd:dateTime	Starttime of object (Required for timeline)	1368-01-01T00:00:00Z 1368-01-01

- Change the following in the Mapping Table. Those names should be exactly how it appears in the preview Table above. Double-check white spaces, capital letters, commas, dots, etc, when specifying the mapping. You can read the detailed instructions of the mapping later
- Note the statistic row should be filled
- This mapping is simplified for the tutorial. You can map as much as you wish

Your Column Header	Target Column Header
year	Item
jan	Object
year	Starttime
jan	Statistic (Timeline)

Step 4B: (Only with statistics) Customise your chart

Title of Statistics	Chart Type	Chart Point Style	Chart Point Size	
Statistics	✓ Line chart / Scatter Plot / Bar Chart	Circle ⌄	10	1)Specify the title of statistics, used for lengend name of Timeline (default: "Statistics"). Select 2) a chart type, and 3) a point style of the chart, and 4) a point size of the chart in pixel

- Step 4B allows you to cutstomise the design of the chart to display the weather data. You can change the Title of Statistics as "Jan Mean Temp Vienna", while Chart Type is "line chart" and chart Point Style "square", and Chart Point Size "10".

- Scroll Down to Step 5

Perform Comparison Search (Step 5)

Step 5: Comparison Search: search entities in Rekisstory

- You can compare the history of up to four entities
- Checkbox "Include inverse relations" gives more comprehensive results about the entity, such as art works for a person entity, but will slow down the proc
- (Only with statistics) Checkbox "Merge with statistics" allows you to merge the specific entity with statistics (INNER JOIN), by assinging it as the unique id Mapping Table will be then overwritten. If you do not check, the statistics are added without merging (OUTER JOIN). Only one out of four entities can be
- It would take **5 - 60 seconds** to see the results (without inverse relations)

Language for your search results (default: English)

`Start typing here to find a language`

1st entity with Q-name from Wikidata (Q9202 Statue of Liberty, Q154708 Shinji Kagawa, Q219831 The Night Watch)

`Napoleo` ❗ Include inverse relations ☐ Merge with statistics ☐ Value on Y-axis (for stats mixing) `0`

Q78994:Napoleonic Wars 1803–1815 wars involving the French Empire
Q517:Napoleon French military leader, French Emperor 1804–1814 and again in 1815
Q71084:First French Empire empire of Napoleon I of France (1804–1815)
Q152499:Peninsular War 1807–1814 conflict during the Napoleonic Wars
Q7721:Napoleon III President and Emperor of the French (1808–1873)
Q223936:Kingdom of Italy kingdom in southern Europe between 1805 and 1814
Q7723:Napoleon II Emperor of the French (1815)

statistics ☐ Value on Y-axis (for stats mixing) `Type in integer`

statistics ☐ Value on Y-axis (for stats mixing) `Type in integer`

statistics ☐ Value on Y-axis (for stats mixing) `Type in integer`

- Here is the comparison search of our service
- Start typing "Napoleon" in the 1st mandatory input field, and select the right one "Napoleon Bonaparte (Q517)". Make sure "Q517" remains in the input box (not "Napoleon" as text). Type "15" for Value on Y-axis (for stats mixing) for this entity.
- Start typing "Haydn" in the 2nd input field, and select the right one "Josef Haydn (Q7349)". Type "20" for Value on Y-axis (for stats mixing) for this entity.
- Start typing "Wolfgang Amadeus Mozart" in the 3rd input field, and select the right one (Q254). Type "25" for Value on Y-axis (for stats mixing) for this entity.
- Start typing "Beethoven" in the 4th input field, and select the right one "Ludwig van Beethoven (Q255)". Type "30" for Value on Y-axis (for stats mixing) for this entity.
- If you have a better idea, change the entities as you wish, or try it later
- Leave other input as is
- Check you input all again, and click Search & Integrate Data and wait for tens of seconds (i.e. query would take a couple of minutes without progress bar)

The result

Search Results for

Napoleon   **Joseph Haydn**   **Wolfgang Amadeus Mozart**   **Ludwig van Beethoven**   **year**

●   ●   ●   ●   ●
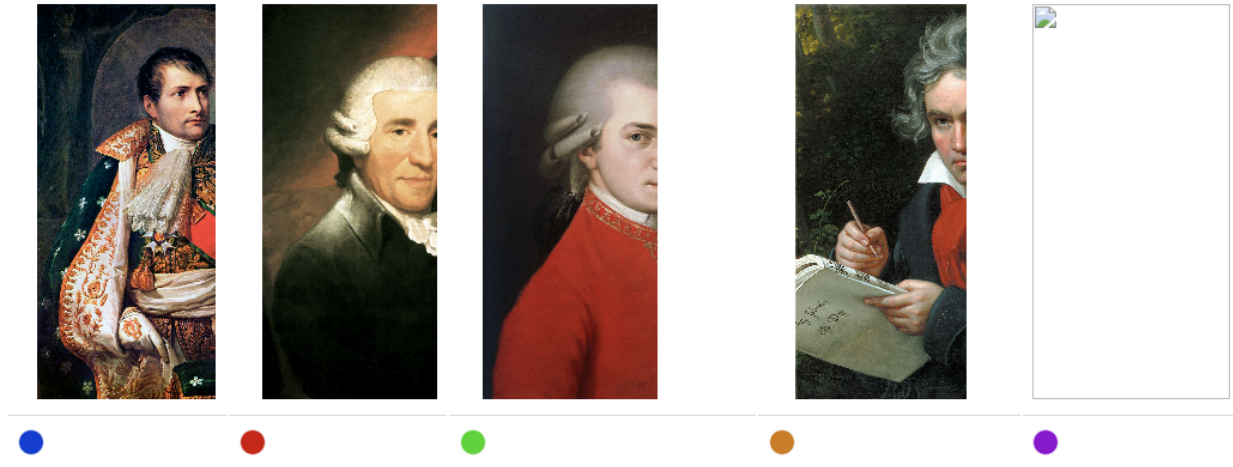
- First, you will see images of the entities specified. Year is empty, because we did not specify the mapping
- Scroll down

122.508 seconds

If table, timeline, and map presentation look strange, check the fact in the table too. They may be due to the irregularities of the source data (e.g. age mis-calculation may occur especially when there are multiple dates or data absence). It seems Wikidata has discrepancy between the date in RDF (raw data) and in HTML (displayed label)

Filter your results with a keyword

Type a keyword...

C... ⇕	Entity ⇕	Rel... ⇕	Ob... ⇕	Inverse Rel... ⇕	Image ⇕	S... ⇕	Start... ⇕	E... ⇕	End... ⇕	Data Source ⇕
●	Joseph Haydn	place of birth	Rohrau			1732-03-31	0 years old (during lifetime)	No date	? years old (during lifetime)	https://www.wikidata.org/
●	Wolfgang Amadeus Mozart	place of birth	Salzburg			1756-01-27	0 years old (during lifetime)	No date	? years old (during lifetime)	https://www.wikidata.org/
●	Napoleon	residence	Ajaccio			1769-01-01	0 years old (before lifetime)	1779-01-01	9 years old (before lifetime)	https://www.wikidata.org/

- If you see the table like above (pin-colors may look different). Your data integration is a success. Examine the table to check if the data is correct and your mapping is what is expected
- Pin-color of each entity should be identical for Table and Map view (Timeline view is different)
- Clock Next button at the right bottom of the table to move the result pages. If you are able to spot the years of measurements in the Starttime column and the temperatures in the Object column, everything is correct. By default, the table is sorted by startime, but you can change the order of results by clicking a header

- You can also filter your results by typing a keyword input box above the table

🟣	year		-0,2			1782-01-01	None years old (None)	None	None years old (None)	Externally imported data
🟢	Wolfgang Amadeus Mozart	spouse	Constanze Mozart			1782-08-04	26 years old (during lifetime)	1791-12-05	35 years old (during lifetime)	https://www.wikidata.org/
🟣	year		2,0			1783-01-01	None years old (None)	None	None years old (None)	Externally imported data
🔵	Napoleon	educated at	École Militaire			1784-01-01	14 years old (during lifetime)	No date	? years old (during lifetime)	https://www.wikidata.org/

Showing **11** to **20** of **232** results   Previous   1   **2**   3   ...   24   Next

- If you click Download data in CSV, you can also download the query results

🔵	Napoleon	place of burial	Valley of the Tomb			No date
🔵	Napoleon	place of burial	Valley of the Tomb			No date

Showing **161** to **164** of **164** results

Download data in CSV
Download context data in CSV

- Scroll down to see the map view. Would you be able to see the map below?
- What do you think of the distribution of information about the entities? You can clearly see the association with France, Germany, and Austria where Napoleon, Beethoven, Haydn and Mozart were born or lived respectively
- Click plots to see details about it and click hyperlinks to check the data
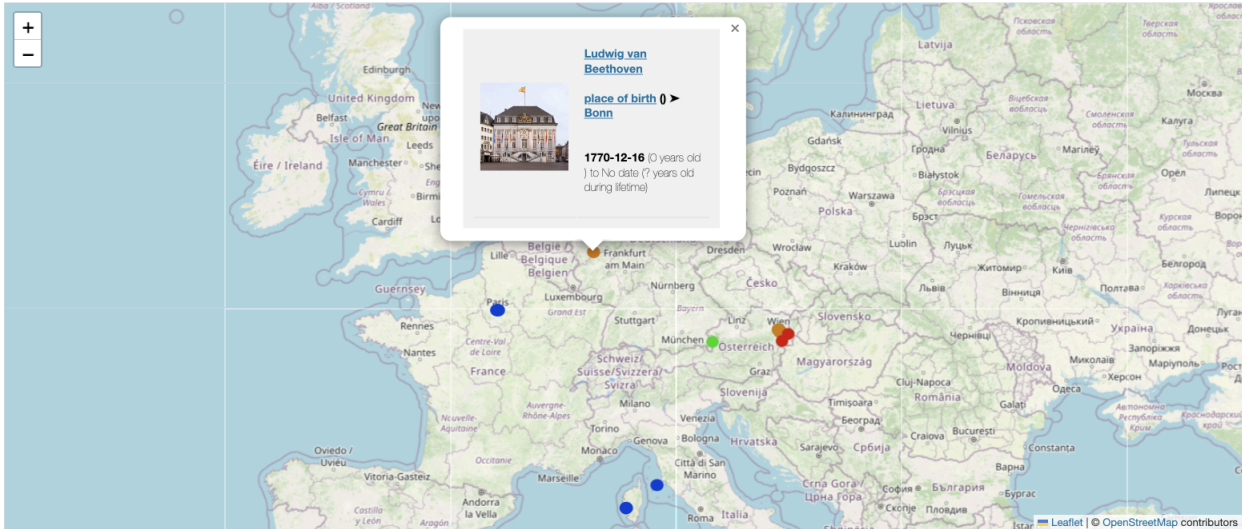
- Scroll down to see the statistics in the timeline view. It looks a little messy at the first glance, but if you zoom in, you can see the data more clearly.
- Notice that your configurations of chart style, point style, and point size are applied
- In the example below, the red line and square points are the January mean temperatures. You can verify it in the legend at the bottom left of the Timeline. While the line and all points on line15 show timeline of Napoleon (blue), line 20 for Haydn (orange), line 25 for Mozart (purple), line 30 for Beethoven (green)
- In this way, you can compare five types of data easily and interactively
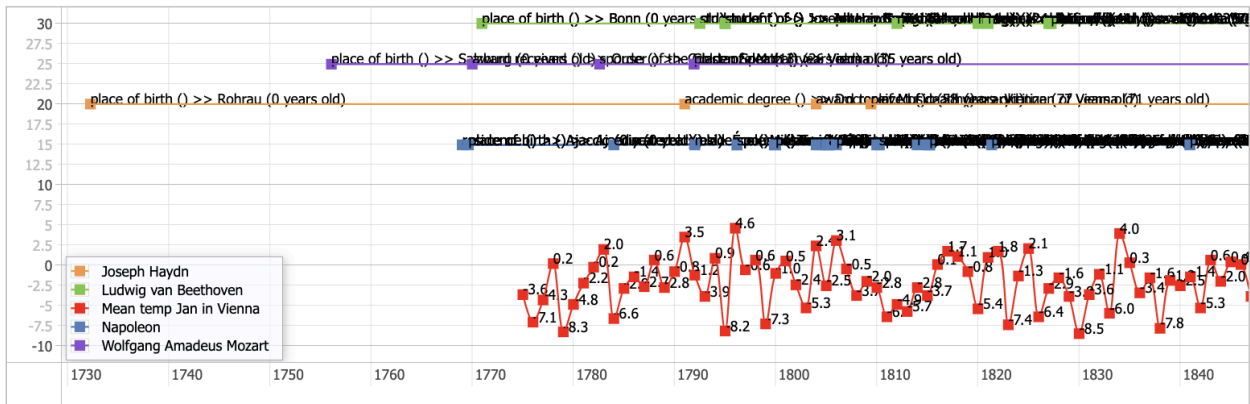
Adding Statistics or Merging Statistics

- In an example above, we added the temperature data to the Comparison Search results. This is handy when you would like to keep them separate. But you may rather want to "merge" them. Merging means you consolidate the statistics with the search result of one entity. This function assigns the user specified entity identifier and name to all records of the statistics.
- How to do that? Follow the same step of data import in the above example until the last step of the data integration. Before you click "Search & Integrate Data", you specify with which entity you would like to be merged. You can find checkboxes next to entity input

fields. Check the box next to the entity of "Napoleon Bonaparte (Q517)". Note you can only select one entity to merge with the statistics.

Step 5: Comparison Search: search entities in Rekisstory

- You can compare the history of up to four entities
- Checkbox "Include inverse relations" gives more comprehensive results about the entity, such as art works for a person entity, but will slow down the pr
- (Only with statistics) Checkbox "Merge with statistics" allows you to merge the specific entity with statistics (INNER JOIN), by assinging it as the unique setting of the entity in the Mapping Table will be then overwritten. If you do not check, the statistics are added without merging (OUTER JOIN). Only or
- It would take **5 - 60 seconds** to see the results (without inverse relations)

Language for your search results (default: English)
`Start typing here to find a language`

1st entity with Q-name from Wikidata (Q9202 Statue of Liberty, Q154708 Shinji Kagawa, Q219831 The Night Watch)
`Q517`   ❗ Include inverse relations ☐ Merge with statistics ☑ Value on Y-axis (for stats mixing) `0`

- The results would look like below

🔵	Napoleon	place of birth	Ajaccio			1769-08-15	0 years old (during lifetime)	No date	? years old (during lifetime)	https://www.wikidata.org/
🔵	Napoleon	place of birth	Ajaccio			1769-08-15	0 years old (during lifetime)	No date	? years old (during lifetime)	https://www.wikidata.org/
🔴	Napoleon		-3,6			1775-01-01	None years old (None)	None	None years old (None)	Externally imported data
🔴	Napoleon		-7,1			1776-01-01	None years old (None)	None	None years old (None)	Externally imported data
🔴	Napoleon		-4,3			1777-01-01	None years old (None)	None	None years old (None)	Externally imported data

- As you can see, the statistics are now labeled as Napoleon. This is different from completely independent rows like you saw before. You have "mapped" Napoleon to the statistics
- If you download CSV file, you can also see the consolidated data as such
- In both adding and merging, our data integration is a kind of "full outer join" in the sense of SQL query, because the multiple sources do not interfere with each other, when they are joined. However, we do not join in the SQL sense, because we do not consolidate table rows by the foreign key. For merging, we "provisionally" assign a foreign key to the importing data in order for the users to be able to relate them to the specified entity. In other words, we do not merge two records from two different sources by aligning them through a foreign key. For example, we still see two distinct records/rows, even if the values of the foreign key (entity ID) and the dates from the two data sources are exactly the same.

Conclusion

As you know, this is not for a serious research topic. It is just a quick demonstration of data integration. However, it is hoped that this tutorial and Rekkistory stimulated your inspiration. You can find more interesting data and proper analysis for your purpose. So, be bold to mix a broad range of data with our tool and make a serendipitous discovery. Who knows you might become a new ground-breaking historian! Thank you!