

The background of the entire image is a complex, abstract geometric pattern. It consists of numerous small, irregular shapes, primarily squares and triangles, in a variety of colors including red, orange, yellow, green, blue, and grey. The pattern is dense and covers the entire frame, creating a textured, mosaic-like effect.

**PEGGY  
GUGGENHEIM  
COLLECTION**

**Study of the  
online collection**

Giovanna Pichierri



# Scenario

The **Peggy Guggenheim Collection** is one of the most important museums of European and North American art of the twentieth century in Italy. It is located in Peggy Guggenheim's former home, Palazzo Venier dei Leoni, on the Grand Canal in Venice.

The **website-collection** is the result of the collaborative efforts of Basilico and the museum staff, whose experience and skills contributed to the conception and development of the project. The website has been devised to better respond to the needs of its users, and its architecture and its 'user journey' are specifically designed for visitors.



# Objectives

---

## PEGGY GUGGENHEIM COLLECTION

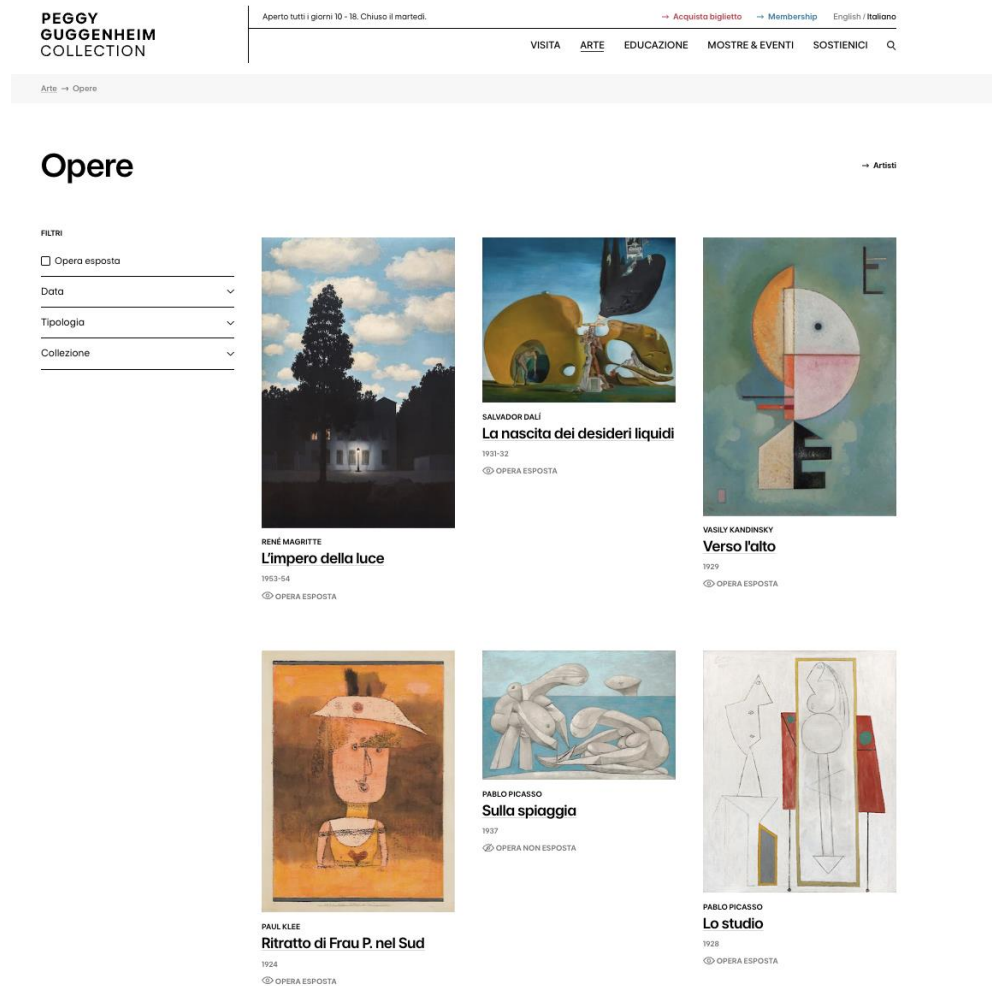
Crate the data-base of the Peggy Guggenheim Collection of Venice for study:

- How have the works been catalogued?
- Are there inconsistencies in the cataloguing method?
- How many artworks are/are not exhibited?
- Which artists are most represented in the collection?
- Which historical periods are included?



# Phase 1: selection of sources

To create a database with the informations from the online collection, I scraped data from the website:



<https://www.guggenheim-venice.it/it/arte/opere/>

# Phase 2: website inspection

Searching for elements in the HTML code:



VASILY KANDINSKY  
**Verso l'alto**  
1929  
OPERA ESPOSTA

`<p class="ArtworkPreview-artist">Vasily Kandinsky</p>`  
`><h1 class="ArtworkPreview-title">...</h1>`  
`<p class="ArtworkPreview-date">1929</p>`  
`><p class="ArtworkPreview-status">`

```
# Rileva tutti i campi
artisti = driver.find_elements(By.CLASS_NAME, "ArtworkPreview-artist")
titoli = driver.find_elements(By.CLASS_NAME, "ArtworkPreview-title")
date = driver.find_elements(By.CLASS_NAME, "ArtworkPreview-date")
stati = driver.find_elements(By.CLASS_NAME, "ArtworkPreview-status")
```

# Phase 3: scraping

```
def estrai_dati_opere(driver):
    artisti = driver.find_elements(By.CLASS_NAME, "ArtworkPreview-artist")
    titoli = driver.find_elements(By.CLASS_NAME, "ArtworkPreview-title")
    date = driver.find_elements(By.CLASS_NAME, "ArtworkPreview-date")
    stati = driver.find_elements(By.CLASS_NAME, "ArtworkPreview-status")

    num_opere = min(len(artisti), len(titoli), len(date), len(stati))
    print(f"Trovate {num_opere} opere in questa pagina.")

    opere = []
    for i in range(num_opere):
        artista = artisti[i].text.strip()
        titolo = titoli[i].text.strip()
        anno = date[i].text.strip()
        stato = stati[i].text.strip()

        opere.append({
            'Artista': artista,
            'Titolo': titolo,
            'Anno': anno,
            'Stato': stato
        })
        print(f" {i+1}: {artista} - {titolo} - {anno} - {stato}")
    return opere

def salva_csv(opere, filename):
    with open(filename, 'w', encoding='utf-8', newline='') as f:
        writer = csv.DictWriter(f, fieldnames=['Artista', 'Titolo', 'Anno', 'Stato'])
        writer.writeheader()
        for opera in opere:
            writer.writerow(opera)
    print(f"Tutti i dati salvati in '{filename}'")
```



# Phase 3: scraping

```
Apro la pagina: https://www.guggenheim-venice.it/it/arte/opere/
Pagina caricata correttamente.
Scroll 1/30...
Scroll 2/30...
Fine del caricamento: altezza pagina non cambia più.
Trovate 24 opere (basato sul minimo tra i campi)
1: Artista: RENÉ MAGRITTE | Titolo: L'impero della luce | Anno: 1953-54 | Stato: OPERA ESPOSTA
2: Artista: SALVADOR DALÍ | Titolo: La nascita dei desideri liquidi | Anno: 1931-32 | Stato: OPERA ESPOSTA
3: Artista: VASILY KANDINSKY | Titolo: Verso l'alto | Anno: 1929 | Stato: OPERA ESPOSTA
4: Artista: PAUL KLEE | Titolo: Ritratto di Frau P. nel Sud | Anno: 1924 | Stato: OPERA ESPOSTA
5: Artista: PABLO PICASSO | Titolo: Sulla spiaggia | Anno: 1937 | Stato: OPERA NON ESPOSTA
6: Artista: PABLO PICASSO | Titolo: Lo studio | Anno: 1928 | Stato: OPERA ESPOSTA
7: Artista: JACKSON POLLOCK | Titolo: La donna luna | Anno: 1942 | Stato: OPERA ESPOSTA
8: Artista: JACKSON POLLOCK | Titolo: Foresta incantata | Anno: 1947 | Stato: OPERA ESPOSTA
9: Artista: GIORGIO DE CHIRICO | Titolo: La torre rossa | Anno: 1913 | Stato: OPERA NON ESPOSTA
10: Artista: MAX ERNST | Titolo: La vestizione della sposa | Anno: 1940 | Stato: OPERA ESPOSTA
11: Artista: CONSTANTIN BRANCUSI | Titolo: Maiastra | Anno: 1912 C. | Stato: OPERA ESPOSTA
12: Artista: KAZIMIR MALEVICH | Titolo: Senza titolo | Anno: 1916 C. | Stato: OPERA ESPOSTA
13: Artista: GINO SEVERINI | Titolo: Mare=Ballerina | Anno: 1914 | Stato: OPERA ESPOSTA
14: Artista: JEAN (HANS) ARP | Titolo: Scarpa azzurra rovesciata con due tacchi sotto una volta nera | Anno: 1925 C. | Stato: OPERA NON ESPOSTA
15: Artista: ALBERTO GIACOMETTI | Titolo: Donna sgozzata | Anno: 1932 | Stato: OPERA NON ESPOSTA
16: Artista: ARTISTA NON RICONOSCIUTO SALAMPASU | Titolo: Maschera ("mukinka") | Anno: PRIMA METÀ DEL XX SECOLO | Stato: OPERA NON ESPOSTA
17: Artista: BERENICE ABBOTT | Titolo: Peggy Guggenheim | Anno: 1926 C. | Stato: OPERA NON ESPOSTA
18: Artista: BERENICE ABBOTT | Titolo: Galleria surrealista, Art of This Century | Anno: 1942 | Stato: OPERA ESPOSTA
19: Artista: BERENICE ABBOTT | Titolo: Galleria astratta, Art of This Century | Anno: 1942 | Stato: OPERA ESPOSTA
...
22: Artista: BERENICE ABBOTT | Titolo: Galleria a luce naturale, Art of This Century | Anno: 1942 | Stato: OPERA NON ESPOSTA
23: Artista: BERENICE ABBOTT | Titolo: Galleria a luce naturale, Art of This Century | Anno: 1942 | Stato: OPERA NON ESPOSTA
24: Artista: ARTISTA NON RICONOSCIUTO ABELAM O BOIKEN | Titolo: Elemento di una casa cerimoniale | Anno: METÀ DEL XX SECOLO | Stato: OPERA NON ESPOSTA
Dati salvati in 'guggenheim_opere_completo.csv'
```

```
base_url = "https://www.guggenheim-venice.it/it/arte/opere/"
# Pagine da 1 (index base url) a 26
pagina_inizio = 1
pagina_fine = 26
output_csv = "guggenheim_opere_tutte_le_pagine.csv"
```

During the scraping process, only the data displayed on the first page was initially extracted. To fix this issue, I added a section in the code that specifies the first and last pages to be scraped.



# Phase 3: scraping

Problem: Selenium code extracts only 2 artworks

Solution: implementing infinite scrolling and adding wait times

```
Trovate 24 opere (basato sul minimo comune tra campi)
1: Artista: RENÉ MAGRITTE | Titolo: L'impero della luce | Anno: 1953-54 | Stato: OPERA ESPOSTA
2: Artista: SALVADOR DALÍ | Titolo: La nascita dei desideri liquidi | Anno: 1931-32 | Stato: OPERA ESPOSTA
3: Artista: | Titolo: | Anno: | Stato:
4: Artista: | Titolo: | Anno: | Stato:
5: Artista: | Titolo: | Anno: | Stato:
6: Artista: | Titolo: | Anno: | Stato:
7: Artista: | Titolo: | Anno: | Stato:
8: Artista: | Titolo: | Anno: | Stato:
```

*The fact that the Selenium code extracts only 2 artworks from the Peggy Guggenheim Venice website probably indicates that:*

*The artwork data is loaded dynamically using pagination or on-demand loading mechanisms.*

```
def scroll_infinito(driver, pausa=3, max_scroll=30):
    """Scorre la pagina verso il basso per caricare tutti i contenuti dinamici."""
    last_height = driver.execute_script("return document.body.scrollHeight")
    for i in range(max_scroll):
        print(f"Scroll {i+1}/{max_scroll}...")
        driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")
        time.sleep(pausa) # attesa per caricamento contenuti

        new_height = driver.execute_script("return document.body.scrollHeight")
        if new_height == last_height:
            print("Fine del caricamento: altezza pagina non cambia più.")
            break
        last_height = new_height
```



# Phase 4: CSV output

```
# Salvataggio CSV
with open(output_csv, 'w', encoding='utf-8', newline='') as f:
    writer = csv.DictWriter(f, fieldnames=['Artista', 'Titolo', 'Anno', 'Stato'])
    writer.writeheader()
    for opera in opere:
        writer.writerow(opera)
```

CSV file export. In tabular format, with information divided into columns.

guggenheim\_venezia\_opere\_tutte\_le\_pagine

	Titolo	Anno	Stato
RENE MAGRITTE	L'impero della luce	1953-54	OPERA ESPOSTA
SALVADOR DALÍ	La nascita dei desideri liquidi	1931-32	OPERA ESPOSTA
VASILY KANDINSKY	Verso l'alto	1929	OPERA ESPOSTA
PAUL KLEE	Ritratto di Frau P. nel Sud	1924	OPERA ESPOSTA
PABLO PICASSO	Sulla spiaggia	1937	OPERA NON ESPOSTA
PABLO PICASSO	Lo studio	1928	OPERA ESPOSTA
JACKSON POLLOCK	La donna luna	1942	OPERA ESPOSTA
JACKSON POLLOCK	Foresta incantata	1947	OPERA ESPOSTA
GIORGIO DE CHIRICO	La torre rossa	1913	OPERA NON ESPOSTA
MAX ERNST	La vestizione della sposa	1940	OPERA ESPOSTA
CONSTANTIN BRANCUSI	Maiastra	1912 C.	OPERA ESPOSTA
KAZIMIR MALEVICH	Senza titolo	1916 C.	OPERA ESPOSTA
GINO SEVERINI	Mare-Ballerina	1914	OPERA ESPOSTA
JEAN (HANS) ARP	Scarpa azzurra rovesciata con due tacchi sotto una volta nera	1925 C.	OPERA NON ESPOSTA
ALBERTO GIACOMETTI	Donna sgozzata	1932	OPERA NON ESPOSTA
ARTISTA NON RICONOSCIUTO SALAMPASU	Maschera ("mukinka")	PRIMA METÀ DEL XX SECOLO	OPERA NON ESPOSTA
BERENICE ABBOTT	Peggy Guggenheim	1926 C.	OPERA NON ESPOSTA
BERENICE ABBOTT	Galleria surrealista, Art of This Century	1942	OPERA ESPOSTA
BERENICE ABBOTT	Galleria astratta, Art of This Century	1942	OPERA ESPOSTA

# Phase 5: ETL – import Open Refine

**OpenRefine** A power tool for working with messy data. Do you want to be notified of new OpenRefine releases and events? [Yes](#) [No](#) ([privacy info](#))

Create project « start over Configure parsing options Project name  Tags  Create project »

Open project  
Import project  
Language settings  
Extensions

	Artista	Titolo	Anno	Stato
1.	RENÉ MAGRITTE	L'impero della luce	1953-54	OPERA ESPOSTA
2.	SALVADOR DALÍ	La nascita dei desideri liquidi	1931-32	OPERA ESPOSTA
3.	VASILY KANDINSKY	Verso l'alto	1929	OPERA ESPOSTA
4.	PAUL KLEE	Ritratto di Frau P. nel Sud	1924	OPERA ESPOSTA
5.	PABLO PICASSO	Sulla spiaggia	1937	OPERA NON ESPOSTA
6.	PABLO PICASSO	Lo studio	1928	OPERA ESPOSTA
7.	JACKSON POLLOCK	La donna luna	1942	OPERA ESPOSTA
8.	JACKSON POLLOCK	Foresta incantata	1947	OPERA ESPOSTA
9.	GIORGIO DE CHIRICO	La torre rossa	1913	OPERA NON ESPOSTA
10.	MAX ERNST	La vestizione della sposa	1940	OPERA ESPOSTA
11.	CONSTANTIN BRANCUSI	Maiestra	1912 C.	OPERA ESPOSTA
12.	KAZIMIR MALEVICH	Senza titolo	1916 C.	OPERA ESPOSTA
13.	GINO SEVERINI	Mare=Ballerina	1914	OPERA ESPOSTA
14.	JEAN (HANS) ARP	Scarpa azzurra rovesciata con due tacchi sotto una volta nera	1925 C.	OPERA NON ESPOSTA
15.	ALBERTO GIACOMETTI	Donna sgozzata	1932	OPERA NON ESPOSTA
16.	ARTISTA NON RICONOSCIUTO SALAMPASU	Maschera ("mukinka")	PRIMA METÀ DEL XX SECOLO	OPERA NON ESPOSTA
17.	BERENICE ABBOTT	Peggy Guggenheim	1926 C.	OPERA NON ESPOSTA

**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  
RDF/XML files

Columns are separated by  
☐ commas (CSV)  
☐ tabs (TSV)  
☒ custom ;

☒ Use character " " to enclose cells containing column separators  
☐ Trim leading & trailing whitespace from strings  
Escape special characters with \

☒ Ignore first 1 line(s) at beginning of file  
☒ Parse next 1 line(s) as column headers  
☐ Column names (comma separated)

☐ Discard initial 0 row(s) of data  
☐ Load at most 0 row(s) of data

☐ Attempt to parse cell text into numbers  
☒ Store blank rows  
☒ Store blank columns  
☒ Store blank cells as nulls  
☐ Store file source  
☐ Store archive file

Disable auto preview

Version 3.9.3 [TRUNK]  
Preferences  
Help  
About

Creating the OpenRefine project by importing the CSV file with the correct settings for display.

# Phase 5: ETL – ‘Artista’ column

OpenRefine guggenheim venezia

Facet / Filter Undo / Redo 0 / 0

Refresh Reset all Remove all

**Artista** change

22 choices Sort by: name count Cluster

ARTISTA GIAPPONESE NON RICONOSCIUTO 1

ARTISTA NON RICONOSCIUTO 2

ARTISTA NON RICONOSCIUTO ABELAM O BOIKEN 1

ARTISTA NON RICONOSCIUTO BAGA 1

ARTISTA NON RICONOSCIUTO BAMANA 2

ARTISTA NON RICONOSCIUTO CHAMBRI 1

ARTISTA NON RICONOSCIUTO CHIMÚ (REGNO DI CHIMOR) 2

ARTISTA NON RICONOSCIUTO CUBEO 1

ARTISTA NON RICONOSCIUTO DOGON 3

ARTISTA NON RICONOSCIUTO IATMUL OCCIDENTALE 1

ARTISTA NON RICONOSCIUTO KOTA 1

ARTISTA NON RICONOSCIUTO MADAK 1

ARTISTA NON RICONOSCIUTO MANDARA (O TABAR) 1

ARTISTA NON RICONOSCIUTO



**Artista** change

1 choice Sort by: name count Cluster

ARTISTA NON RICONOSCIUTO 36 edit include

Facet by choice counts

36 artworks in the collection do not have clear indications regarding the author.

During cataloguing, it was decided to add a detail related to the period or type of artwork to the label ‘ARTISTA NON RICONOSCIUTO’.

However, this information is not necessary for the purposes of my study. I have decided to create a single label for all artworks in the collection where we do not have clear information about the artist.



# Phase 5: ETL – ‘Anno’ column

Anno		change	invert	reset
98 choices		Sort by: name		count
Cluster				
2005	1			
2006	2			
2007	1			
2017	1			
300 AEV–400 EV	3			
900–1470 EV	2			
FINE XIX SECOLO – INIZIO XX SECOLO	1			
FINE XIX–INIZIO XX SECOLO	1			
INIZIO DEL XX SECOLO	3			
INIZIO DEL XX SECOLO	1			
PRIMA METÀ XX SECOLO	2			
S.D.	11			
XVI–INIZIO XX SECOLO	1			
Facet by choice counts				

Very often, artworks do not have precise indications about the year of creation, or often the creation lasted several years.

For this reason, during cataloguing, the Anno (Year) column has been filled in different ways:

(yyyy) ?

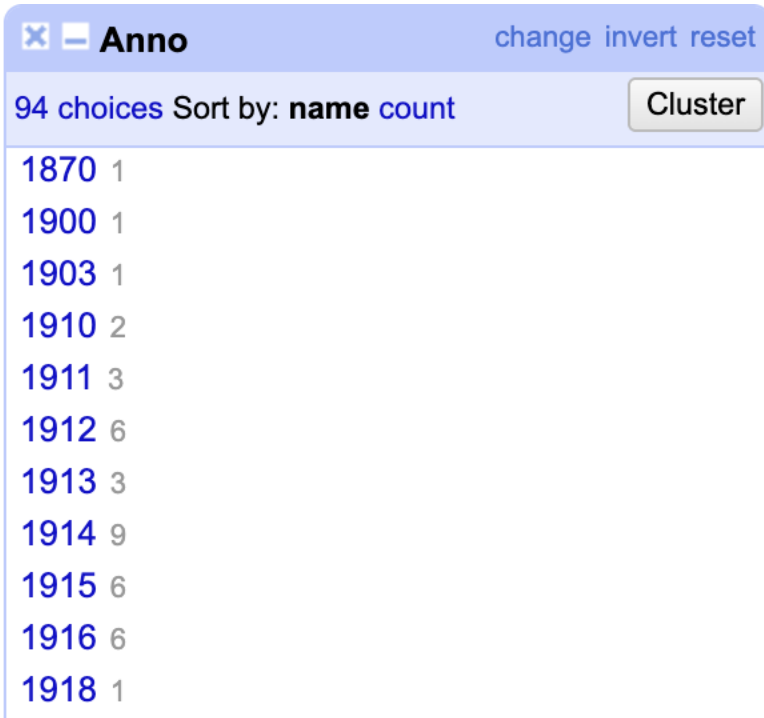
Aaxx/aayy

Vague century indications

First half/second half

S.D.

# Phase 5: ETL – ‘Anno’ column



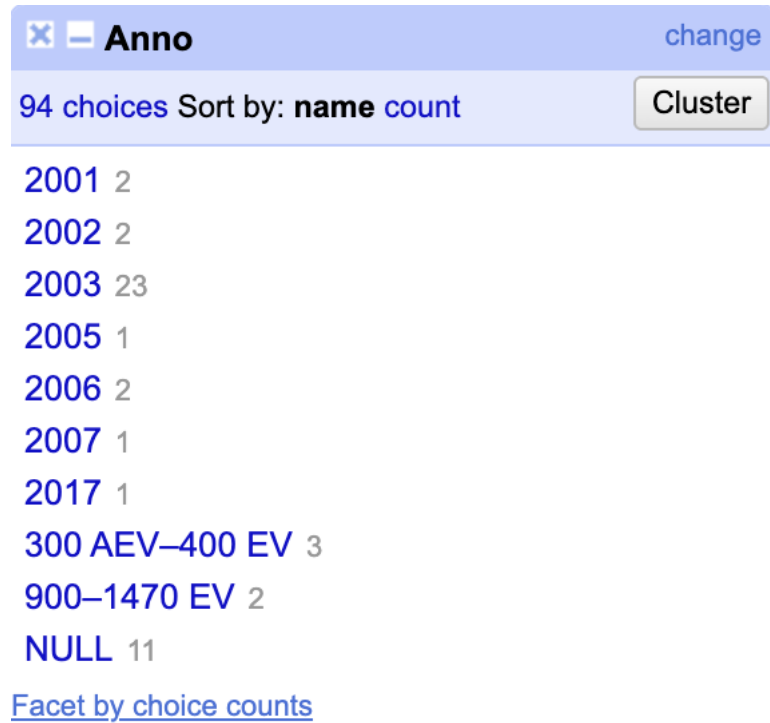
Anno	
1870	1
1900	1
1903	1
1910	2
1911	3
1912	6
1913	3
1914	9
1915	6
1916	6
1918	1

For the purposes of my study, I wanted to have clear information about the temporality of the artworks in the collection.

Therefore, I decided to make some changes to the Year column:

- I converted the data from string to number.
- I modified the entries where there was a '?' or 'C.' (circa), removing that information to obtain a number.
- For date ranges, I chose to keep the most recent year, presumed to be the end (e.g., 1914-1915 → 1915).
- For year indications, I converted Roman numerals to Arabic numerals.

# Phase 5: ETL – ‘Anno’ column



The screenshot shows a data visualization interface for the 'Anno' column. At the top, there is a header bar with a close button (X), a minus sign, the label 'Anno', and a 'change' link. Below the header, it says '94 choices' and 'Sort by: name count'. There is a 'Cluster' button. The main area displays a list of date ranges and their counts: 2001 (2), 2002 (2), 2003 (23), 2005 (1), 2006 (2), 2007 (1), 2017 (1), 300 AEV–400 EV (3), 900–1470 EV (2), and NULL (11). At the bottom, there is a link 'Facet by choice counts'.

Year/Range	Count
2001	2
2002	2
2003	23
2005	1
2006	2
2007	1
2017	1
300 AEV–400 EV	3
900–1470 EV	2
NULL	11

Two cases in which I did not modify the cell to a number:  
S.D. (No Date) → NULL

3 artworks where AEV (Before Common Era) was indicated

2 artworks that don't have a date to be included in the analysis

I decided to proceed this way because the artworks with these characteristics are only 5, so they will not affect my analyses.



## Phase 5: ETL – ‘data\_info’ column

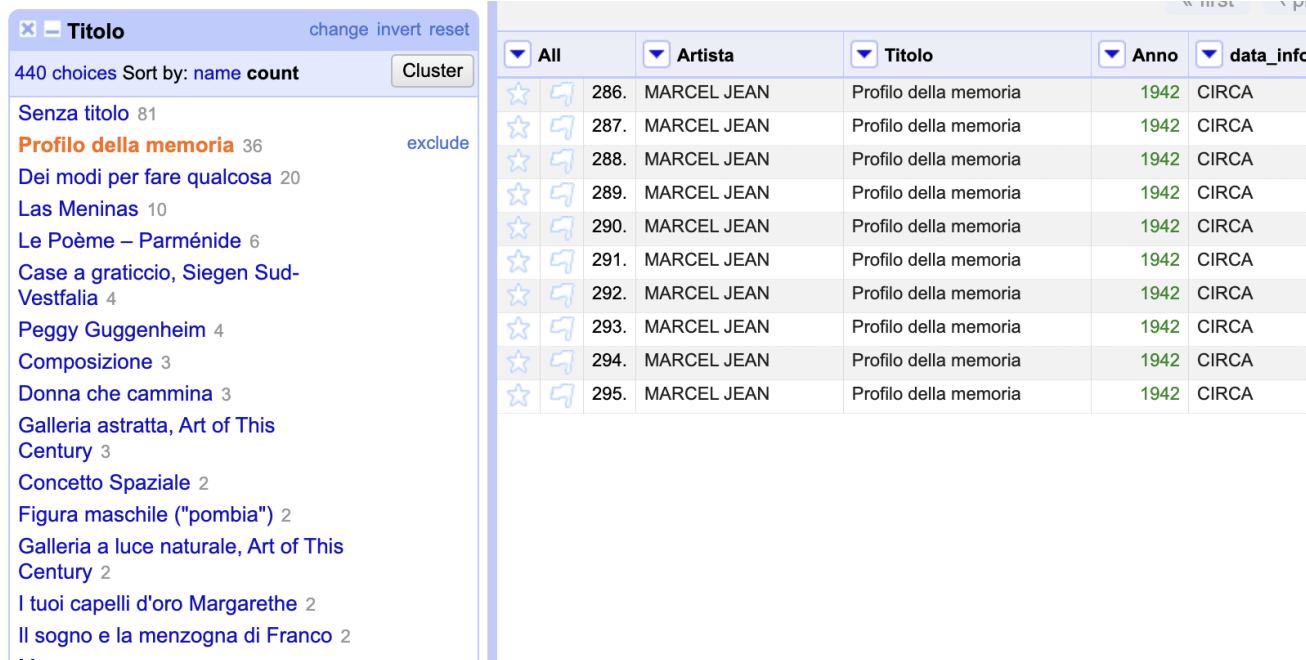
▼ Anno	▼ data_info
1954	CIRCA
1932	CIRCA
1929	TRUE
1924	TRUE
1937	TRUE
1928	TRUE
1942	TRUE
1947	TRUE
1913	TRUE
1940	TRUE

Having manipulated the Anno' column, I decided to add the column *data\_info* in order not to completely lose the previous information:

TRUE: no modifications are made; the date is exactly as indicated during cataloguing.

CIRCA: indicates that a modifications are made for analysis purposes (modifications described in the previous slides).

# Phase 5: ETL – duplicates



The screenshot shows the Peggy Guggenheim Collection database interface. On the left, a sidebar titled "Titolo" displays a list of 440 choices, sorted by name count. The list includes titles like "Senza titolo", "Profilo della memoria", "Dei modi per fare qualcosa", "Las Meninas", "Le Poème – Parménide", "Case a graticcio, Siegen Sud-Vestfalia", "Peggy Guggenheim", "Composizione", "Donna che cammina", "Galleria astratta, Art of This Century", "Concetto Spaziale", "Figura maschile ('pombia')", "Galleria a luce naturale, Art of This Century", "I tuoi capelli d'oro Margarethe", and "Il sogno e la menzogna di Franco". A "Cluster" button is visible. The main area displays a table of artworks, with columns for "All", "Artista", "Titolo", "Anno", and "data\_info". The table shows a list of 295 records, with the first 10 records being duplicates of the same artwork: "Profilo della memoria" by Marcel Jean, dated 1942, and categorized as "CIRCA".

All	Artista	Titolo	Anno	data_info
☆	286. MARCEL JEAN	Profilo della memoria	1942	CIRCA
☆	287. MARCEL JEAN	Profilo della memoria	1942	CIRCA
☆	288. MARCEL JEAN	Profilo della memoria	1942	CIRCA
☆	289. MARCEL JEAN	Profilo della memoria	1942	CIRCA
☆	290. MARCEL JEAN	Profilo della memoria	1942	CIRCA
☆	291. MARCEL JEAN	Profilo della memoria	1942	CIRCA
☆	292. MARCEL JEAN	Profilo della memoria	1942	CIRCA
☆	293. MARCEL JEAN	Profilo della memoria	1942	CIRCA
☆	294. MARCEL JEAN	Profilo della memoria	1942	CIRCA
☆	295. MARCEL JEAN	Profilo della memoria	1942	CIRCA

More than 50 records had duplicate values.

These artworks were catalogued with the same information for Artist, Title, and Year.

# Phase 5: ETL – duplicates



MARCEL JEAN  
**Profilo della memoria**  
1935-42  
OPERA NON ESPOSTA



MARCEL JEAN  
**Profilo della memoria**  
1935-42  
OPERA NON ESPOSTA



MARCEL JEAN  
**Profilo della memoria**  
1935-42  
OPERA NON ESPOSTA



MARCEL JEAN  
**Profilo della memoria**  
1935-42  
OPERA NON ESPOSTA



MARCEL JEAN  
**Profilo della memoria**  
1935-42  
OPERA NON ESPOSTA



MARCEL JEAN  
**Profilo della memoria**  
1935-42  
OPERA NON ESPOSTA



MARCEL JEAN  
**Profilo della memoria**  
1935-42  
OPERA NON ESPOSTA



MARCEL JEAN  
**Profilo della memoria**  
1935-42  
OPERA NON ESPOSTA

In artistic contexts, this is not necessarily an error.

Very often, artworks have the same references because they are part of the same series.

This is also very common in lithographic works or photographs.

Duplicate records in this context can be useful for my study to understand how many series of artworks are part of the collection.



# Phase 6: Conclusion



The online catalog of the Peggy Guggenheim Collection in Venice is overall well done.

During the scraping process, I did not encounter any major blocking issues. Regarding data completeness, it is 100% complete. I did not find any empty cells.

The most significant problems are related to **date** entries: different methods were used to express the same information, maybe because the work was not always done by the same person or the cataloging came from multiple sources.

Additionally, for artworks belonging to the same series, which during the ETL phase may appear as **duplicates**, the handling could have been more thorough, for example, by including a serial number alongside the title. Now, the database is ready to answer to more specific questions about the collection's content.

## Phase 6: Conclusion

Now, the database is ready to answers more specific questions about the collection's content and setting.

To be continued....

