**Materials**

*Digitisation and annotation -* The archival data located at the Ghent Opera were acquired utilizing a mobile and user-friendly equipment called ScanTent. This device operates in conjunction with the Docscan application on a mobile phone, enabling high-resolution scans to be captured under optimal lighting conditions. Since at the start of the project I lacked the skills to automatically crop the images to size, the majority of the images were cropped manually. However the last collections were cropped automatically, the code that was used for this can be found in the notebook *cropper.ipynb*. Since there were some errors in this, some of these images were cropped manually nonetheless. Due to the size of the corpus, the original annotations were initially relatively simple (see fig. 1). All the collections between 1893 and 1934 were digitised and annotated which yielded over 10,000 images. It should be noted that there is no data between 1914-1918 and 1921-1922 since the Antwerp opera was closed in those periods.

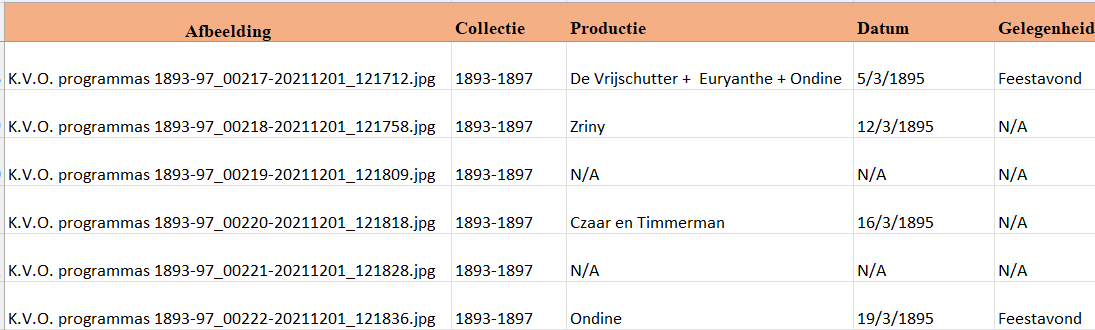


Figure : excerpt of the original annotations.

The original dataframe only consisted of five columns; the *Afbeelding* column, reflected the filename of the image of the corresponding annotations. The *Collectie* indicated in which bundle the programme leaflet was found. On most occasions this corresponded with operatic season of the performance, which runs approximately from October to April. Some seasons were however bundled together into one collection, as is the case with collection 1893-1897 in the example above. The *Productie* indicates the title of the performed work, following the spelling of the programme leaflet. If it was part of a multiple bill, i.e. more than one production performed at the same occasion, the different production are separated by a ‘+’ . The multiple performance are annotated in the order that they were performed on the evening. In some cases, one image contained information on several performances, without it being a multiple bill. In this case, a new row was created where the metadata was stored separately, but the *afbeelding*-column would stay the same. The *Datum* column indicates the date on which the performance took place formatted as dd/mm/yyyy. Lastly, the *Gelegenheid* column indicated the specific occasion for which the performance took place such as a gala-evening, a party or popular display (*volksvertoning*) . In most cases this was N/A.

*Pre-processing –* The pre-processing steps can be followed in the notebook *DF\_Cleaner.* The annotations were loaded in a jupyter notebook called *DF\_cleaner* in which the following pre-processing steps were conducted. For each column trailing and leading whitespaces which might have occurred during the annotation process, were removed. A new, Boolean column was created indicated whether the programme leaflet advertised a multiple bill (True) or not (False). There were 526 multiple bills identified. Another column called ‘Performance’ was created, which gave a random 6-digit ID to each row. In order to make the individual works of the multiple bills countable, each row that contained a ‘+’ was split to create individual rows for each performance. Through the Performance\_ID it remained clear that these productions were performed together. This step is illustrated in figures 2 and 3.

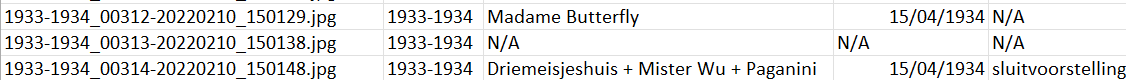


Figure : Dataframe before splitting the multiple bills.

Dataframe after adding performance\_ID and splitting multiple bills

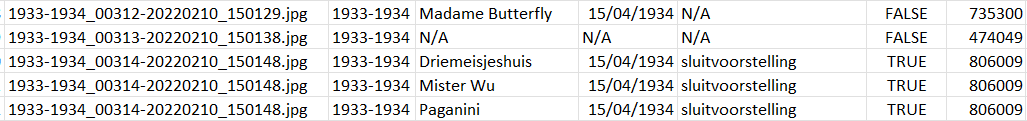


Figure : Dataframe after adding performance\_ID and splitting multiple bills.

There existed spelling discrepancies in the titling of the productions. Specifically, there were instances in which the Dutch title was utilized, while at other times, the original title was employed. Additionally, there was notable variation within the Dutch spelling. For instance, the original Dutch work *Quinten Massijs*, was invariably written as *Quinten Massys, Quinten Metsys, Quinten-Massijs, Quinten-Massys* or *Quinten Massijs.* In order to make sure that these productions were interpreted as the same work, a new column with the normalised spelling was created. When it was unclear what the original spelling was, the most common way of spelling was adopted as the normalized one. This was done using a self-established dictionary in which the normalized spelling was the key, and the spelling variations the values. An excerpt of this dictionary can be found in fig. 4. Since this dictionary with spelling variations was created manually, it is possible that there are some irregularities that went under the radar.

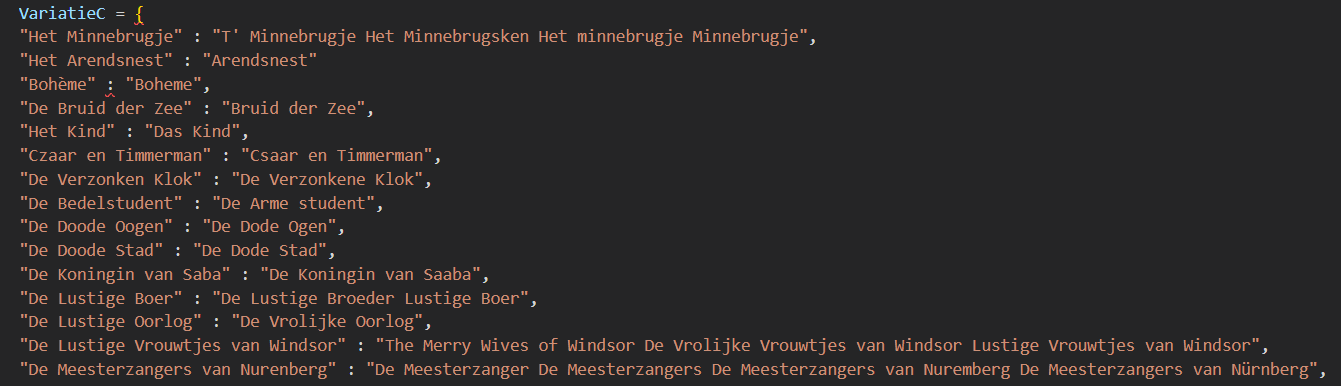
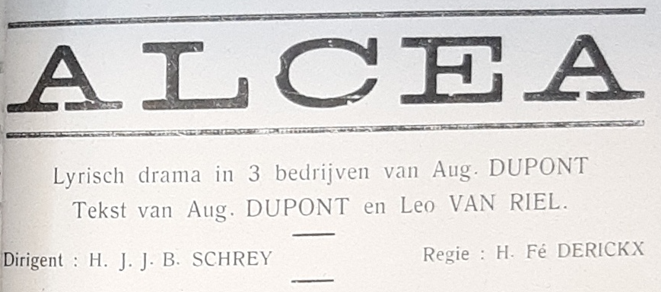


Figure : Dictionary used to map and normalise spelling variations in the dataframe?

Furthermore, the date column was converted to datetime datatype. A substantial proportion of the images in the dataset contained no information on the repertoire and were either blank (as a verso side of a programme leaflet) or contained advertisements. Since these will not be used in the analyses, they were removed from the dataframe. Lastly, duplicates were also removed. These pre-processing steps resulted in the reduction of the dataframe from 10992 rows (one row per image) to 4470 rows (one row per performance).

*Additional metadata –* for each unique work in the dataset, additional metadata was gathered in a text file. This was done by creating a dictionary where the key was the normalised spelling of the work, and the value was a list consisting of five items:

* Original language e.g. DUI, NL, FR …
* Composer e.g. Richard Wagner, Albert Lortzing, Franz Léhar…
* The year of the works’ first performance.
* The original title e.g. Das Rheingold, Der Bettelstudent, Les Templiers …
* The genre e.g. opera buffa, lyrisch drama, operette…

The International Music Score Library Project (IMSLP) was used as reference to gather this material. The IMSLP however, contains no metadata on the operatic genre of a work, this data was predominantly gathered via ChatGPT. A sample of ChatGPT’s responses was checked to see if the output was adequate. Since this was the case, it was used to gather the rest of the genres, although it is possible that there are still some errors in this. In some cases and especially with local opera’s, ChatGPT was unable to give a response. In those cases I used the genre that was advertised on the programme leaflet, as in the example blow (see fig. ??)

After exporting the dataframe one additional column was added manually to indicate the director(s) of the Antwerp opera house at the time of the performance. Below are the first fifteen rows of the dataframe as it was used for further analyses.

