Deliverable 3 (WP2):

Report to Accompany "One corpus of Text"

European Union HORIZON-MSCA-2022-SE-01-01

Project 101129889 — PortADa

Authoring Institution and Responsible Beneficiary:

UNIVERSITAT DE BARCELONA

Project Name	PortADa. "Port Arrivals Data. Automatic data collection for a large-scale comparative history of 19th century shipping: a Digital Humanities approach to maritime heritage"
Grant	European Union, HORIZON-MSCA-2022-SE-01-01
Project Number	Project 101129889 — PortADa
Document Title	One corpus of text in txt file format
Responsible Beneficiary	Universitat de Barcelona
Deliverable, Work Package	D3: One corpus of text in txt file format (WP 2)
Type	Written report
Date	31 March 2025
Number of pages	7 (including cover)
File Name	101129889_PortADa_D3_One_corpus_of_text_in_txt_file_format
Authors	Authored by Jordi Ibarz (Project Coordinator, U. Barcelona) Reviewed by Brendan J. von Briesen (Manager Assistant; U. Barcelona)
Contact	brendan.vonBriesen@ub.edu

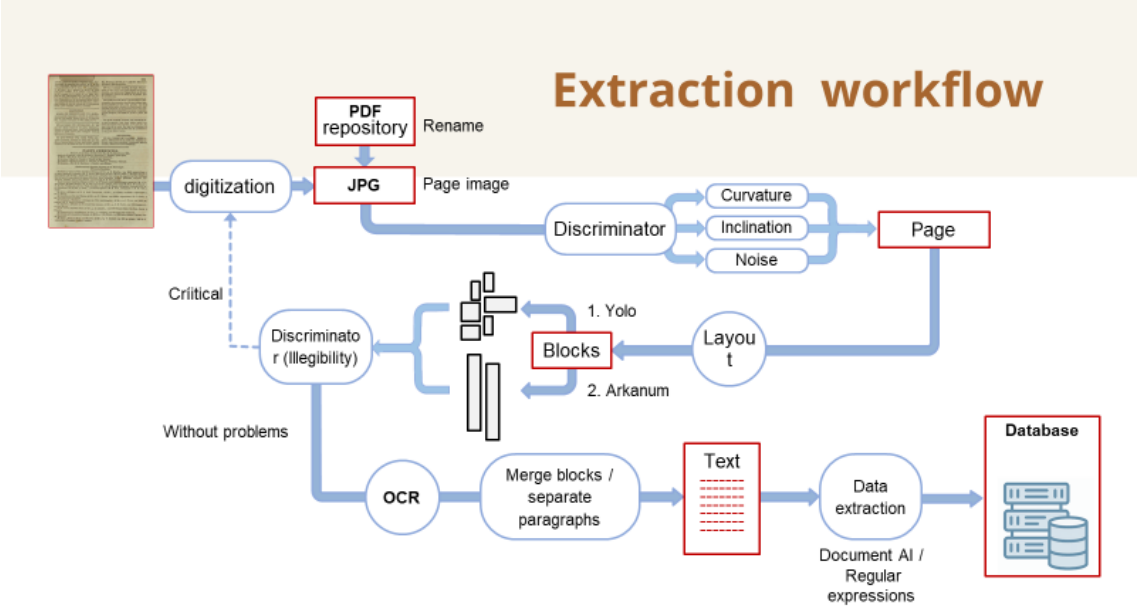
Funded by the European Union

"*Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.*"

1.- Process of extracting the TXTs

Each local node is responsible for obtaining the information corresponding to its port.

We have followed similar procedures to obtain these documents. The basic scheme for obtaining this information is shown here.



We work with a server, using a software called PApiCli (PortADa Api Client), which was developed by the computer engineers of the project. The software can be downloaded here:

https://drive.google.com/file/d/1ZS8J87v2nZhC_UXabAlrF9B8ZpWgY7ED/view

Through this application, the historians of the project have processed the images of the newspapers of Barcelona, Buenos Aires, Havana and Marseille.

The material we have prepared corresponds to the text related to news about the arrival of ships at the port.

2. Directory Structure Where the Material is Located

The material is located in a Google Drive directory: Portada/Entrega_txt

Accessible via this link:

https://drive.google.com/drive/folders/1XG9vzTcscrM7fTJJq4gvk6ZApEZqQdQu?usp=sharing

It is organized into various folders, one for each port, and within each folder, there is another folder where the texts are stored. These are organized into subfolders by year.

The structure is as follows:

Entrega_txt

- Barcelona
  - TXTs
    - 1850_OCR (1850_01_01_BCN_DB_U_10_000....)
    - 1851_OCR
    - 1852_OCR
    - ...
- Buenos Aires
  - TXTs
    - 1850_txt (1850_01_07_BUE_GM_U_04_000,,,)
    - 1851_txt
    - 1852_txt
- Habana
  - TXTs
    - 1850 (1850_01_01_HAB_DM_U_04_0_000...)
    - 1851
    - 1852
- Marseille
  - TXTs
    - 1850 (1850_01_01_MAR_SM_U_03_000_000...)
    - 1851
    - 1852

3. Document Names

The names of the documents are standardized and respond to the following criteria:

- The file name includes 6 distinct concepts: publication date, port, newspaper name, edition, news page, and block number.

- The name does not have to contain all the concepts, only those that are essential for uniquely identifying each file.
- The length of the name is invariably 28 characters, including the separators. This means 20 characters of content and 8 separators ("_").

Example: 1854_04_25_BUE_LP_0_08_000

Publication Date: [YYYY_MM_DD] 1854_04_25_

They are always in this format because this order allows us to organize the documentation chronologically.

Port: BUE_

Possible values (BCN, BUE, HAB, MAR), are always three digits

Newspaper Name: LP_

The example uses the acronym of *La Prensa*.

Possible values (DB *Diario de Barcelona*, DM *Diario de la Marina*; SM *le Semaphonre de Marsellie*, etc.) These are always two digits.

Edition: One digit

Possible values (M morning, T afternoon, N night, U unique). In those cases where it is always a unique edition, this must be made explicit in the name with the value U. It is always a digit and should not be encoded numerically.

To avoid problems, any type of confusion, and subsequent renaming work, we should always use these values in Spanish.

News Page: 08_ two digits

Observation: Sometimes the pagination is consecutive for each year or even extends over several years. We usually take the page number as the position of the page within the digital document if it is the entire newspaper, or generally the position within each issue of the newspaper, e.g., page 8 of the copy where the news is found.

Block Number: three digits

It is a numerical value. Possible values (from 000 to 999). This indicator refers to specific needs of the project. In our work strategy, at a certain point in the process, we have chosen to divide the pages into blocks. Since there are very large pages with a lot of information, this can lead to many blocks, hence the existence of the three digits.

When we are working with the pages in a phase of the work process, this part of the name is not included, so the name of the document (page) is 4 digits shorter.

4. Summary of material delivered

A summary of all the texts existing in each of these directories is shown in the following table:

Year	Barcelona	Buenos Aires	Havana	Marsella
1850	1627	455	2711	42019
1851	1606	389	3430	3508
1852	1698	289	4527	3532
1853	1941	616	2770	3558
1854	1256	837	3835	3474
1855	2540	493	2943	2923
1856	2663	586	3415	3521
1857	2416	473	5595	3655
1858	2636	673	1980	3714
1859	2575	851	1883	4068
1860	2275	723	1424	3843
1861	2354	799	2336	3657
1862	2265	839	2801	2955
1863	2559	821	2589	4192
1864	2422	791	2445	3605
1865	2349	723	1878	3417
1866	2621	681	2481	3865
1867	2742	709	2095	4083
1868	2435	677	2138	3236
1869	2600	624	2078	4127
1870	1716	635	2042	24627
1871	2697	635	2423	4176
1872	1736	948	2680	48089
1873	2404	945	2346	5244
1874	2755		2365	4854
1875	2291	736	2846	3426
1876	2458	282	3036	2834
1877	2558		2465	2295
1878	2808		3349	5545
1879	2788	655	2819	5441
1880	2683	791	2432	5409
1881	2996	646	2875	4143
1882	3601	865	2920	5191
1883	3246	597	na	5410
1884	2953	828	2701	4842
1885	1774	916	na	5236
1886	2193	239	4738	4115
1887	1853	505	6625	4442
1888	2488	260	5971	4419

1889	1772	26	6706	4273
1890	2132	415	6817	3984
1891	2180	597	6525	2884
1892	1895	707	6573	3596
1893	2202	674	4086	2194
1894	1721	789	5436	2426
1895	1663	698	6745	3067
1896	1838	573	3030	3720
1897	2262	555	2623	4319
1898	2081	593	na	4763
1899	1892	740	1275	4810
1900	1715	1042	218	4584
1901	2317	1032	3782	3946
1902	2230	1009	2678	3360
1903	2160	992	4167	2664
1904	2346	1034	3715	3137
1905	2307	1007	3679	2773
1906	2059	812	1563	3390
1907	1977	1072	1415	3609
1908	1772	730	na	2786
1909	2237	1196	1611	3105
1910	2295	976	830	4574
1911	1955	992	2210	
1912	1418	1069	2681	3463
1913	1644	1041	1733	3114
1914	1531	1019	1168	
1915	1373		1127	
Total	146552	44922	192380	345231

As shown in the table, the acquisition of the texts from which we need to extract information is practically complete.

For Barcelona, we have 100% of the necessary texts. An error has been detected for 1854, the year for which only the first page containing the arrivals of ships at the ports has been processed. When the news spans two or three pages, we have only considered the first one.

For Buenos Aires, we are only missing the extraction for 1874, 1877, and 1878, as well as for 1915. The absence of these early years is due to the newspapers being in PDF format and falling outside the workflow of that local node. This material is pending processing. Additionally, there are dates for which we do not have copies in either PDF or paper in the newspaper archive. Furthermore, during the OCR processing with PAPICLI, approximately 3,200 images with OCR processing errors were found, and they are currently being reprocessed. Out of an approximate total of 24,000 days, about 1,850 copies are missing, for which we are exploring alternatives in other newspaper archives. Of those 1,850, approximately 350 correspond to dates when the newspaper was not published, as we have verified

from the edition numbering. In about 750 cases, the copy existed but did not record the corresponding notes. Just over 4% of the total expected remains pending.

For Havana, about 4 years are still pending: 1883, 1885, 1898, and 1908. For these years we do not have a digital copy of *Diario de la Marina*. This is indicated in the table with an "na" (not available). Cuban historians held a working meeting with the Economic Society of Friends of the Country of Cuba, which also preserves a collection of *Diario de la Marina*. They were able to confirm that the missing newspapers are in a delicate state of preservation, making their reproduction impossible. We are trying to locate paper copies in other archives. If that is not possible, we will look for alternative newspapers. In the case of Havana, processing with Paplici has only been carried out for the 1850s, during the Cuban team's stay at the Summer School in Barcelona. When they tried to process the remaining information from Cuba, the server did not work, not even through VPN. For this reason, processing has been done with free software, Tesseract. The results are not of sufficient quality, so we will reprocess the remaining material using the planned workflow for the other nodes. We will carry out the processing from European territory to resolve the mentioned issues.

For Marseille, all years have been processed, except for 1911 and 1914-1915, which are now processing. To create the layout, pages could be processed in blocks or by columns. The general criterion has been to process by columns. Only in a few years for Marseille—1850, 1870, and 1872—has the process been done in blocks, which explains the difference in the number of TXT files shown in the general summary.

In all cases, through various automatic, semi-automatic, or manual procedures, we have selected those pages of the newspapers where the news of ship arrivals at the ports appeared. In the case of Barcelona, news of arrivals at the ports usually occupies several pages, so what is in the corresponding directories refers to the complete texts of the pages containing that type of news. In some cases, the entire page refers to ship arrivals, but in others, there is different information on the same page. In contrast, for Buenos Aires, since the process of obtaining information was different and only the part of the newspaper referring to ship arrivals was digitized, what is digitized is only the news, and the texts correspond exclusively to the material we are interested in. For Marseille and Havana, the news is also inserted in pages with other information. In these last two cases, since the pages are very large, there is a lot of other information in the TXTs that is different from the ship arrivals at the ports.

5. Conclusion

We can consider this phase of the project for obtaining the news TXTs completed. The few years that remain to be completed are not significant compared to the total work done, and they will also be processed in the coming weeks.

Based on the material already available, we have started the next phase, which is the extraction of information for the construction of a database.