Deliverable 1.1 (WP1):

Data Management Plan

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	Plan for Data Management
Responsible Beneficiary	Universitat de Barcelona
Deliverable, Work Package, and Task	Deliverable 1.1; Work Package 1; Task 1.2
Type	Written report: Data Management Plan
Date	28 June 2024
Number of pages	7 (including cover)
File Name	101129889_PortADa_DMP
Authors	Universitat de Barcelona (UB)
Collaborator	Brendan von Briesen (Project Staff)
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.- Executive Summary

This Data Management Plan covers the creation and management of data for the PortADa project. The project uses digital images of publicly available historical newspapers that are over one hundred years old. All data and metadata will be managed ethically, securely, and in accordance with the FAIR principles during the project and for at least ten years after completion.

2.- Data Overview

One of the two main objectives of the project is the creation of a database regarding the arrivals of ships into the ports of Barcelona, Marseille, Havana, and Buenos Aires between 1850 and 1914.

An initial estimate calculates approximately 1.6 million ship arrivals as the data to be generated. This will require working with around 90,000 pages of daily historical press from the different ports. The information contained in each entry is very similar across the different ports, although not identical. In general, it includes the arrival date, voyage duration, port of origin, ports of call, type, name and Gross Registered Tonnage (GRT) of the ship, the name of the captain or master of the vessel, as well as the quantity, unit, and type of goods transported and the merchant to whom they are destined.

The source of this information is images or PDFs that are freely and publicly available for scientific activities, as in our case, found in various historical digital press repositories. We use the metadata from these digital archives. In some cases, due to the absence of these sources, digital capture will need to be done directly using the project's resources, from existing newspapers in libraries and archives. Specifically, the origin of the digitized images and their associated metadata will be as follows:

- ARCA. Arxiu de Revistes Catalanes Antigues. [https://arca.bnc.cat/arcabib_pro/ca/inicio/inicio.do]
- Biblioteca Digital del Caribe. [https://dloc.com/es/title-sets/UF00001565]
- BVPH. Biblioteca Virtual de Prensa Histórica. [https://prensahistorica.mcu.es/es/inicio/inicio.do]
- Hemeroteca digital. Arxiu Històric de la Ciutat de Barcelona. [https://ahcbdigital.bcn.cat/hemeroteca/colleccio/Diario+de+Barcelona]
- Gallica. [https://gallica.bnf.fr/accueil/es/content/accueil-es?mode=desktop]

At the end of the project, various types of data will be generated, mostly textual, which will be formatted to fit into a database.

Except for the metadata from each of the newspapers we will work with, the rest of the data will be newly obtained through the application of OCR on the images used.

The utility of these data can be of interest to the scientific community in general, and specifically to those dedicated to economic history, maritime history, history of technology, social history of labor, among others. It can also be of interest to genealogists and a non-specialized public.

3.- FAIR Data

3.1.- Facilitating the Findability of Data, Including Metadata

The data will be identified using a persistent identifier, specifically a Digital Object Identifier (DOI), at the time of publication.

Enriched metadata will be provided to enable findability. The metadata created will adhere to the standards used for citing and describing databases in DataCite [https://schema.datacite.org/meta/kernel-4.5/].

Keywords will be used in the metadata to optimize subsequent discovery and reuse. Metadata will be offered in a way that can be harvested and indexed.

3.2.- Making Data Accessible

Repository:

The data will be deposited in a trusted repository. Each of the databases generated by the local nodes working on each port will ensure that they are in a trusted repository linked to their research institution. The final or consolidated database will be available in CSUC Data Repositories [https://dataverse.csuc.cat/about.xhtml].

There is a standardized way to deposit data in the aforementioned repository. In any case, this repository guarantees the assignment of an identifier to the data, resolving the identifier to a digital object, DOI.

Data:

All data will be made available to the public.

A 12-month embargo will be applied after the completion of the project to allow time for the publication of the first analyses obtained with the processed data. The enormous volume of data obtained will not be usable until the third year of the project, so it will not be until then that we can begin to work with it, justifying the embargo. Data will be accessible via a free and standardized access protocol.

During the project, each local node will be responsible for the storage and security of the data related to its port. This storage must be done in the institutional repositories of the participating organizations to ensure data security and accessibility for the members of each local node. These participating members will be able to access the data without explicit authorization. It is possible to authorize other academic personnel specialized in maritime history or other related fields to use the data. Local node leaders will be informed 15 days in advance, and in the absence of explicit objections, the data will be made available to the applicants. Usable data will only be available from the third year; previously generated material will be raw, error-laden, and unrefined, intended for internal use only.

Once the project is completed, the general criterion for access to the generated data will be maximum openness.

Metadata:

Metadata will be made publicly available under the CC0 public domain license in accordance with the grant agreement. This metadata will contain information that allows the user to access the data.

The data will remain available and findable for at least 10 years after the project's completion, according to the conditions established in CORA. The same applies to the availability of metadata.

Accessing the data will not require specific software. The software generated for obtaining this data will be open-source and available in the GitHub repository, under a license that permits its use and potential improvement.

3.3. Data Interoperability

To achieve data interoperability, best practices in the field will be followed, such as using standard data formats for data exchange like JSON, XML, and CSV. Descriptive metadata will be included in the datasets. Common vocabularies, controlled word lists, and ontologies will be used or created to ensure consistency and comprehensibility for all users. The basic ontology to be used will be that developed as the SeaLiT Ontology [https://zenodo.org/records/5964240] which will be adapted or completed according to the project's needs.

The generated ontologies or vocabularies will be openly published to allow for reuse, improvement, and expansion.

The project data will include qualified references to other data.

3.4. Increasing Data Reusability

Through the project's website and the documentation accompanying the various databases, the necessary information will be provided to validate the data analysis and facilitate its reuse.

The data will be available for free in the public domain to allow the widest possible reuse.

The data produced in the project can be reused by third parties once the project is completed, and even before if the predetermined conditions are met.

The provenance of the data will be exhaustively documented using appropriate standards.

The work carried out during the first two years of the project will be aimed at ensuring the quality of the data. This means that, as far as academically possible, the automatic extraction of the data will include all records and all information contained in each record. Additionally, a whole year’s work will be specifically dedicated to reviewing and cleaning the data. Data quality will be ensured through the thorough review of significant samples of automatically obtained data, by comparing the utilized information with available statistical and nominal sources, and by establishing a review protocol to resolve cases of synonymy or nominative equivalence.

4. Other Research Outcomes

Other products generated by the research, such as software, workflows, or protocols, will benefit from the same access conditions as the other data obtained in the project.

There will be two main ways to make these results publicly available. Firstly, through dissemination by presenting at conferences or publishing in academic journals. Secondly, by including them on the project's website or in open-access digital repositories available at the respective universities of the project participants.

5. Resource Allocation

The costs of making the data and other research outcomes FAIR are partly covered for free by CORA, the entity mentioned earlier. Additionally, the project has allocated part of its budget to cover these expenses.

The principal investigator of the project will be responsible for data management. They may require advice and support from the data access committee, which will include one person responsible for this at each local node or a delegated person. This committee will be established during the project development.

The project includes analyzing its future viability once the four years dedicated to it are over. Efforts will be made to secure additional resources or involve an entity linked to the maritime and port world to ensure the data's preservation and future viability.

6. Data Security

While being developed or worked on, the data will be stored in the institutional clouds of each local node under the conditions previously established in this document. Once available for academic and public use, the database and generated web programs will be stored in external Virtual Private Server. Backup copies of all this material will be maintained in the repositories of the University of Barcelona (and any participating universities or research entities that request it), provided they have a system that ensures periodic backups. The database is not considered to include sensitive data that requires special protection.

7. Ethics

There are no ethical or legal issues that could impact data sharing. In any case, it will be ensured that all material produced by each local node is fully available to them at all times and that they have copies of the part of the database they have developed, with the same final quality as the content in the general database.

The generation and use of the databases will comply with the legal and ethical requirements of both the European Union and the country where each database is developed.

8. Conclusions

This Data Management Plan for the PortADa project ensures that the creation and management of data related to the historical ship entries in Barcelona, Marseille, Havana, and Buenos Aires between 1850 and 1914 will adhere to ethical standards, security protocols, and the FAIR principles. The project will generate a comprehensive database with approximately 1.6 million ship entries, extracted from around 90,000 pages of historical newspapers. All data will be made publicly accessible after a maximum of 12-month embargo, stored in trusted repositories, and managed using open-access protocols to facilitate reuse by the scientific community, genealogists, and the general public.

Interoperability will be ensured through the use of standard data formats and ontologies, and data quality will be rigorously reviewed and validated. Additionally, other research outcomes, such as software and protocols, will be freely available. Resource allocation and data security measures are in place to support the long-term sustainability and accessibility of the data, with no identified ethical or legal issues impacting data sharing. The project aims to secure additional resources post-completion to ensure the preservation and future viability of the data.