1. Data Summary

The project will generate two different kinds of data: raw data and processed data.

**Raw data (aka primary data):** archives which have been digitised but not processed nor analysed. This is the source material. It is intended to remain private but may be accessible on demand to other researchers as part of thematic collaborations.

All raw data are pictures of original documents taken in the archives, in accordance with local regulation. For each archive centre visited, there is a folder of pictures which contains documents of direct interest to the project, classified in the same manner as within the institution. All raw data are under jpg format, unless very few exceptions in pdf.

Within the frame of SHIPPAN, some existing raw data are re-used from previous research. It mostly consists in digitised archives from France generated by the Fellow and kept privately; or digitised and online archives from different sources generated by supervisory institutions, and open to the public. These raw data are analysed under a new prism.

It is still difficult to estimate the size of the collected raw data during the project. Based on previous research programme and due to the diversity of the sources, SHIPPAN could generate around 50,000 images. Considering the average size of a digitised document is 2 Mb, the total could reach 100 Gb.

**Processed data (aka secondary data):** documents which summarize and analyse raw data, often made of tabulated and processed raw data from different sources. Processed data are mostly LibreOffice files (ods) or Word files (docx). These data are intended to be public and disseminated as widely as possible to promote the project.

No secondary data will be re-used. All processed data are original and unreleased, generated by the Fellow. The purpose is to cross-reference, compile and analyse the raw data collected to create useful tools for the scientific community as well as to draw relevant conclusions within SHIPPAN’s scope. The main deliverables planned are as follows:

- Inventory of maritime archives related to Health at Sea
- Catalogue of maritime companies (steam long-distance navigation 19th-20th)
- Inventory of International Health rules and regulation (1870-1960)
- Evolving map of sanitary stress at Sea
- Literature review & classified bibliography
- Several articles for publication in peer-reviewed journals

These processed data consist in a dozen of files and could reach 200 Mb in total.

The bibliography, inventories and catalogues will enable researchers who are less expert in the maritime field to find their way easily through the maze of companies and archive centres, and to go straight to the information they need.

For the scientific community in general, processed data will provide compilations of information and unpublished analyses that will enhance a field that is currently little explored. Global, international and transnational studies have been multiplying over the last twenty years or so, and SHIPPAN aims to bring a maritime flavour to it. The data should encourage specialists in social history to take a closer look at maritime issues.

For decision-makers and journalists providing information to the public, these works synthesise complex phenomena from archives scattered around the world. Secondary data will be available for use in sanitary prevention, assessing responsiveness to epidemics and exploring the state of knowledge about health at sea.

2. FAIR data

2.1. Making data findable, including provisions for metadata

Raw data is identified and sometimes made available by the original producer or the custodian organisation. It is not the role of the Fellow to make these data available and/or findable, but rather to guide the public in finding it more easily.

Stored privately and not accessible, raw data will not get a persistent identifier (DOI). These are working documents that have not always been digitised in their entirety or in a clean and professional manner. It will nevertheless be saved in the event that the originals would be destroyed in the archive centres, and pending a comprehensive digitisation policy of these documents.

Depending on the future digitisation policy of the archive centres and on the agreements with them, it could be envisaged to deposit a copy of these digitised archives on Nakala (https://www.nakala.fr/about) according to the protocol explained for processed data below, but it is too early to develop this point at the moment.

Processed data, however, will be made available on Nakala and will therefore be assigned a permanent identifier and metadata, so that it can be standardised, harvested, indexed, and better disseminated.

Nakala requires mandatory metadata for all submissions (data type, title, authors, creation date, licence) and optional metadata (description, keywords, languages, property) which will be systematically filled in. Nakala also gives the possibility of linking a depot with other data published on the platform, which will be used within SHIPPAN but also in relation with other projects.

Literature review & classified bibliography is made available on Zotero (https://www.zotero.org/, a free and open-source reference management software for bibliographic data and related research materials), therefore complying with all requested metadata.

In addition to complying with these expectations, the bibliographic data will comply with the standards of the Bibliothèque Nationale de France and the names of maritime companies and ships will comply with the nomenclatures/entry charts drawn up by the EPCC French Lines & Compagnies, the leading authority on the history of the French merchant navy. These elements are also consistent with another ongoing research project, MarCoMo, in which the Fellow has been involved.
2.2. Making data accessible

Repository:

Nakala is a research data warehouse as well as a platform displaying data for the Humanities and Social Sciences, under the responsibility of the French Huma-Num infrastructure, itself supported by the Centre National de la Recherche Scientifique (CNRS). It is part of the national European Strategy Forum on Research Infrastructures (ESFRI) roadmap, which is in turn aligned with the European Union’s ESFRI framework.

Appropriate arrangements have been explored and SHIPPAN fits perfectly within Nakala’s policies. Thus, the Fellow has already obtained the opening of an account with a capacity of 100 Gb, so he may be able to cope with the volume of raw and processed data together.

A copy of all the processed data linked to Nakala’s permanent identifier will be stored in the SHIPPAN research log (https://steamer.hypotheses.org/). This Hypotheses log is part of the OpenEdition structure, a Humanities and Social sciences publication portal created by the Centre for Open Electronic Publishing, which is regularly harvested and indexed on the Internet and the international network of university libraries.

Data:

With the exception of raw data (see above), all processed data will be made openly and immediately available. There will be no embargo as the creation of these tools will be intellectual property of the Fellow, and therefore protected according to the rules in force.

On Nakala as well as on SteaMER, data are accessible through a free and standardized access protocol. Any persistent identifier (DOI) given to a processed data will be permanent and, in case of update, will remain attached to the document. The former version will stay accessible if necessary. Links that have already been generated will always open the latest version.

If the situation regarding raw data was to change, access to the data would be subject to the rules imposed by each archive centre (which would therefore act as a Data Access Committee) and by the Heritage Codes of the States to which these documents belong. So for the time being, there are no restrictions.

Metadata:

Metadata suggested by Nakala are openly available and licenced under a public domain dedication (CC0). The general and standard data description schema used is Dublin Core.

There is no expiration date on anything stored on Nakala. The Huma-Num research structure (to which Nakala depends) precisely aims to be sustainable and independent. The same goes for metadata which remains even after the possible withdrawal of a document (in the form of a digital tombstone).

With the exception of Zotero (open source) SHIPPAN doesn't use any software. No documentation or reference is therefore necessary.

2.3. Making data interoperable

Interoperability is a crucial element in the reuse of processed SHIPPAN data. Particular attention was therefore paid to the metadata vocabulary, entry charts and formats. Drawing on the experience of other collaborations (Rin MarComo, ANR Portic, BNF) and long-term work with the EPCC French Lines & Companies, interoperability was thus anticipated. The naming standards for ships and shipping companies have been set for a long time by the EPCC and it is therefore these practices which will be applied.

Consequently, the processed data may include qualified references linked to the bibliography or linked to other data, whether from the SHIPPAN program or from another project.
2.4. Increase data re-use

SHIPPAN aims to be a springboard towards other more ambitious research programs, it is therefore necessary to anticipate the re-use of data.

Raw data will be kept in order to be re-used by the Fellow and, if possible and depending on the evolution of the Digital Policies of the Archive Centres, made available to the public. In any case, the exact provenance of each document will be documented (native identification: location, series, and box/file number)

Processed data will be freely available in the public domain to permit the widest re-use as possible. Any third party willing to pursue the research after the Project, to extend, develop, and enrich the data will be able to do so. The quality of the data will be confirmed and certified by a mention at each verification by the SHIPPAN project’s scientific committee.

All data will be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement.

All data uploaded on Nakala and the Research log will benefit from recurrent harvesting and indexations. To facilitate the process, Metadata will be richly described with a plurality of accurate and relevant attributes. In each root directory of Raw and Processed data, a readme file will be provided to explain the original context of the research in which data have been produced. It will contain information and methodology to facilitate an efficient re-using.

3. Other research outputs

The SHIPPAN project generates simple and classic research outputs within the humanities and social sciences: databases (raw and processed), bibliographies, maps, articles, books, oral presentations, and podcasts, etc. In all cases, the digital formats of these productions are subject to exactly the same conditions as the data mentioned above.

Oral presentations and published material, if recorded or printed by a third party, must comply with the rules or the contract of the host organisation or the publisher. The Fellow, for his part, assumes that any external and public intervention, oral or written, concerning his research must be disseminated without restriction. In the case of publications in scientific journals or by academic publishers, priority will be given to those who propose OpenEdition content as quickly as possible.

4. Allocation of resources

To produce the raw data, MSCA project funds were allocated to purchase a handheld scanner. The other costs are inherent to transport to go from one archive center to another.

Thanks to the Huma-Num structure, the use of Nakala (storage and data access) is completely free, whether in direct or indirect costs within the limit of 100 Gb obtained in the initial registration.

To store the raw data and allow permanent offline access, a complete set of all data will be kept on the hard disk of the computer dedicated to the project from the MSCA grant. Also, a backup disk was funded with the same funds.

All data management in SHIPPAN is under the sole responsibility of the Fellow. No external resources are funded to intervene in this process.

As specified above, the retention of data on Nakala is not subject to any time limit. The data will therefore be accessible well after the end of the SHIPPAN project. Nevertheless, it is desirable that the anticipation of interoperability will make it possible to continue to bring this data to life, and to complete it within the framework of another research project.

5. Data security

There is a default storage level which is put into practice as soon as a data is saved in Nakala. Again, there is no time limit once data is stored. The data is described, contextualised and stored
securely within the cutting-edge infrastructures offered by French public organizations dedicated to research.

An advanced level is part of a partnership with the Centre informatique national de l'enseignement supérieur (CINES) and can be considered at the end of the project. In this advanced deposit circuit, long-term preservation is ensured by CINES. CINES provides additional guarantees such as the legal value of the deposit, the integrity of the file, the migration of formats for their readability, etc.

Of course, offline backups will be provided by the Fellow to cover any eventuality (such as a national crash of online resources).

6. Ethics

There are no ethics or legal issues that can have an impact on data sharing.

Raw data are only made available by archive centres if it complies with the laws in force concerning the protection of privacy and the Heritage Codes. The Fellow therefore do not have access to sensitive information that needs to be protected or censored.

Processed data being the result of combined and analysed raw data, the issues are therefore the same.

7. Other issues

Although the European FAIR procedures are very similar, the Fellow is keen to ensure that the data will also be available and compatible with French academic standards and available in United States through the Host Organisation network. In any case, the data will be stored on French or European soil, and priority will be given to structures dedicated to academic research. However, there are currently no plans to use other storage platforms.

On a personal note, the Fellow firmly believes that data mining and artificial intelligence are essential means of disseminating knowledge in the near future. It is therefore preferable for sourced, analysed, and cross-referenced data to be available to everyone in order to avoid and counteract the pollution of false and misleading information.

<table>
<thead>
<tr>
<th>VERSION</th>
<th>PUBLICATION DATE</th>
<th>CHANGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>20.10.2023</td>
<td>Initial version</td>
</tr>
</tbody>
</table>