Téthys: automating a data workflow compiling over 300 years of bathymetric information

Julian Le Deunf, Ronan Jarno, Yann Keramoal, Thierry Schmitt, Morvan Fally, Laurie Biscara, Joseph Dubuis

Téthys: automating a data workflow compiling over 300 years of bathymetric information

Julian Le Deunf
Development bathymetry department
Shom
Brest, France
https://orcid.org/0000-0003-3644-5063

Ronan Jarno
Development bathymetry department
Shom
Brest, France
ronan.jarno@shom.fr

Yann Keramoal
Development bathymetry department
Shom
Brest, France
yann.keramoal@shom.fr

Thierry Schmitt
Development bathymetry department
Shom
Brest, France
thierry.schmitt@shom.fr

Morvan Fally
Development bathymetry department
Shom
Brest, France
morvan.fally@shom.fr

Laurie Biscara
Production bathymetry department
Shom
Brest, France
laurie.biscara@shom.fr

Joseph Dubuis
Development bathymetry department
Shom
Brest, France
joseph.dubuis@shom.fr

***Abstract*—With more than 300 years of existence, the Shom is the oldest active hydrographic service in the world. Compiling and deconflicting that much history is a real challenge. This article will feature the different steps of the workflow allowing Shom to compile over 300 years of bathymetric knowledge. The data access web portal will also be introduced to show the type of user experience built to access this information.**

***Index Terms*—ETL, bathymetric data, quality analysis, data fusion and management**

I. Introduction

The French Hydrographic Service, Shom, provides hydrographic services: nautical charts, sailing directions, list of lights, tide tables and other nautical publications, where applicable, satisfying the needs of safe navigation. These services are compliant with regulation 9 of chapter V of the SOLAS convention (see [1]) by collecting, compiling and disseminating up-to-date hydrographic information on nautical publications, hence ensuring safe navigation. Since 1720, Shom collects hydrographic information on the physical marine environment, particularly bathymetric measurements, allowing the elaboration of nautical products (including nautical charts).

Currently, Shom's bathymetric data are archived in a dedicated Bathymetric Database (BDBS), managed as a stack of potentially overlapping surveys (with over 11 400 surveys). Data from these surveys are derived from different types of sensors: either through mechanical means (lead-line from early 1800 to 1950), or more commonly from acoustic systems (single-beam since the 1950, then using multi-beam sounders since the 80s) or even from optical systems (LiDAR since 2005). Therefore, data held in the BDBS is of varying quality as acquisition procedures have evolved over time. Presently, each cartographic operator generating nautical charts or Digital Terrain Models (DTM) must go through a laborious process of selection of bathymetric information from this database. The fusion of bathymetric data in order to provide a comprehensive compilation is a field of research as shown by numerous publications [2]–[6].

As it stands, this situation constitutes the motivation of the proposal: How to generate, efficiently, the most relevant bathymetric surface from the selection of multiple sources to be used as a reference for most of Shom's uses? To answer this question, the Shom has set up the Téthys project.

The rest of this article is organised as follow, in Section II we briefly introduce the Téthys project, the objectives and the need that this project will respond. In Section III we explain our approach and in particular the data workflow, before presenting in Section IV the result on one tiles with some figures. In Section V we present the web portal to access this deconflicted bathymetric data and to support hydrographic processes. We then discuss in Section VI some conclusions and outline future perspectives raised by this work.

II. The Téthys project

Following a recent effort to digitize all the bathymetric information collected by Shom through the last 300 years, the Téthys project aims at constituting the best current bathymetric knowledge as a bathymetric reference surface, in which a selection of soundings with the best quality (relative to the age, type of sensors used, compliance with existing standards,

...) is being done. Generating this surface will allow Shom to speed up the generation of nautical charts and bathymetric digital elevation models by capitalizing on the selection efforts, along with strengthening the management and the lineage of the original information. This follows the principle: survey once, use many. Generating this reference surface is done through the following process: 1) from the different original soundings, a verification of all the data and metadata content is done, 2) for each of the intersecting data set, the priority (or cut/selection) between overlapping datasets of different qualities and ages is ruled, 3) the compilation of the data is then undertaken following previously defined priorities between datasets. The availability of the reference bathymetric data representing the most appropriate bathymetric knowledge, generated and capitalized upstream by all the actors in a global effort (hydrographers, data managers and operatives) and universally exploited downstream, contributes to the improvement of the exploitation of the Shom's bathymetric data in different use cases: charting and modelling production workflow, survey planification or detection of artefacts or disqualified data on new surveys.

The Téthys project aims at automating data management processes for operational production:

- Make bathymetric knowledge quickly available and therefore quicken the dissemination time of up-to-date information into products to bathymetric data users;
- Improve reference data quality and consistency;
- Enhance Shom's data and products compliance with the security of the mariners;
- Anticipate the inflation of the volume of bathymetric data originating from new sensors and technologies.

Indeed, the Téthys project will provide easy to access bathymetric data for a wide range of actors:

- Survey specifiers will have access to the best knowledge of the current bathymetric information in order to help them planning survey (bathymetric knowledge to be geographically interrogated with respect to strategic (national defense) or commercial criteria (safety of navigation), for example).
- Survey producers will use Téthys data to help them qualifying their new surveys. This qualification will be done by analyzing the new survey with respect to the up-to-date knowledge of existing bathymetric data coverage.
- For the generation of bathymetric product, cartographers or DTM producers will be able to visualize and extract the only data useful for their work areas, with a selection between all the existing data being done and controlled. Shom hopes that Téthys will save at least 10% of the time needed for cartographers to compile the nautical charts.

III. Data management workflow

The challenges on this project are multiple and can be split, on a data centric approach (all information need to be machine readable, scale up with data and IT challenge) and a human approach (need to adopt universal and known deconfliction principles, solve deconfliction problems at the beginning of the process). The large amount of data to be processed requires a robust, high-performance automated process capable of handling massive data. The use of modern ETL (Extract Transform and Load) software along with other technologies (SQL, Python) is particularly adapted to this type of tasks. The adopted workflow is presented on Figure 1.

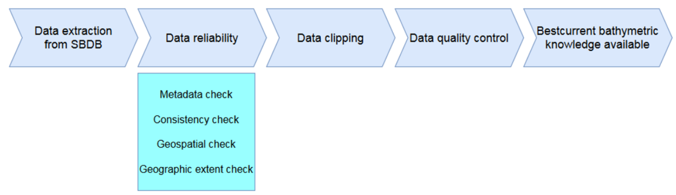

Fig. 1. Téthys workflow overview of each macro-task.

A. Digitalizing effort

Part of the older hydrographic surveys collected by Shom are not fully digitalized. Only some soundings have been selected by the cartographers to be included in the nautical charts. As prerequisite is to be aware of all existing bathymetric knowledge, an important effort has been achieved to make available in numeric format all bathymetric information. This involves operations of scanning and digitization of numerous hydrographic documents and bathymetric sheets, still relevant today, as long as a modern survey will not be realized. Hydrographic survey reports provide also an important source of information, on what the hydrographer discovered during the survey. Without this data transformed into digital format, it is impossible to take advantage of computing capabilities to automate processing steps. The digitization of bathymetric sheets is done with a software developped by Magellium for Shom production, see [7].

B. Data extraction from BDBS

Bathymetric data recorded through time, have been archived in the BDBS. The BDBS is a relational database management system based on PostgreSQL and whose administration and access to the database is operated via Bathy DataBASE Server Administration Tools and BASE Manager respectively. These software packages are developed by the publisher Teledyne CARIS. Records in the database contain descriptive elements (metadatata) relative, both to the methodology of acquisition and actions undertaken during the processing stage. Overall, all these metadata ensure that the archived measures of the seafloor are coherent with the reality with a known level of accuracy and are also in agreement with hydrographic standards that were prevailing at the time the survey has been carried. The extraction from the BDBS aims at providing three components per individual survey at the roots of Téthys workflow:

- Metadata
- Bathymetric coverage polygon
- Bathymetric point cloud

While the metadata and the coverage polygon are readily accessible through PostgreSQL requests, the bathymetric point cloud needs to be downloaded from the database using a dedicated CARIS BDB API, in order to access the information embedded in the csar proprietary file format. The extraction process is dealt in two phases. The first one extracts the metadata and the coverage polygons, which are used first, in order to undertake reliability checks, as described in the next section. Each of these pair of information is then needed to prioritize the data sources in the selection of available surveys (potentially overlapping). The second phase consists of the extraction of the soundings as point cloud. This choice is motivated by the fact that geospatial requests on point cloud are often computer-intensive and hence constitute one of the major bottleneck in the overall process.

C. Data reliability

The extracted data is a mix of surveys collected at different periods and from sensors/methods with heterogeneous qualities. Moreover, not all data sets have the same level of completeness of metadata content. Thus, it is requested to check that metadata of importance to all the subsequent processes are correctly provided. Over 200 individual rules have been implemented. As the rules related to the deconfliction (priority and mutual cooking-cutting of the different source data) are based on intrinsic characteristics of the data set, the aim is to make sure, in a preliminary step, that associated metadata are robust for the deconfliction step (checking for missing fields, previous decision selection/relation with other surveys, errors in sounding and surface clipping…).

1) Metadata check: As BDBS has continuously evolved according to different data models, each metadata is checked to be consistent to the current data model of the database: format, mandatory attributes. Each error is classified according to a typology: some controls are decisive and can block the process of deconfliction, others are considered to be optional as they constitute an aid to decision. The system provides a solution for the database managers to analyze the problem and ask for corrections from experts (hydrographers).

2) Consistency check: In the same data set, each attribute are checked in concordance with the other metadata. For example, vertical and horizontal uncertainties should be compliant with the description of the corresponding attributes from the International Hydrographic Organization S-44 order. This step really helps to ensure the internal consistency of a survey so that future attributes are relevant in the deconfliction process.

3) Geospatial check: Each data set is checked in concordance with the other intersecting data sets in the database. This step deals with logic tests, as metadata allow setting clear priorities. If some attributes are not correctly filled, conflicts between surveys may persist. The concept is to make the deconfliction rules completely machine readable through the attributes. These checks make it possible to verify that the system is able to produce a unique solution, as decided by the hydrographers when they qualify their surveys individually.

4) Geographic extent check: In order to build this reference surface and to achieve the best possible data clipping (deconfliction and compilation) – in the hydrographic sense regarding safety of navigation -, it is particularly important to check the polygons describing the surface of known information. The extent of this polygon is used as a cookie-cutter in the automatic process of selection of the best soundings according to the coverage and describing metadata of the parent survey. Special care must be taken to generate the most realistic envelope of each dataset, called SME – for Surface Minimale Englobante (minimal enclosing surface):

- Unicity: each sounding of the dataset is included in a single SME ;
- Density: soundings which have a distance with the nearest neighbours less than 5 times of the resolution of the dataset are aggregated in the same SME. If not, a new polygon is created, and eventually, the sounding is considered as an isolated sounding if it's impossible to aggregate it with neighbours;
- Representativeness: the contour (internal and external) of the multipolygons is buffered with a distance depending on the characteristics of the survey: horizontal uncertainty and resolution. That is a sensitive point to avoid removing a shoal at the border of the survey with a too loose SME while the area has not been strictly covered.

Importantly, while this polygon simplifies the geometry of the coverage, it is critical that none of the soundings is filtered out through this process, as each measure might need to be displayed on the chart and might be critical for navigation. Figure 2 presents the overall workflow used to compute the multipolygon representing the convex envelope of a bathymetric point cloud. The algorithm takes as input a point cloud (xyz) file and the horizontal uncertainty associated with the soundings (originating from the use of the metadata associated to the survey).

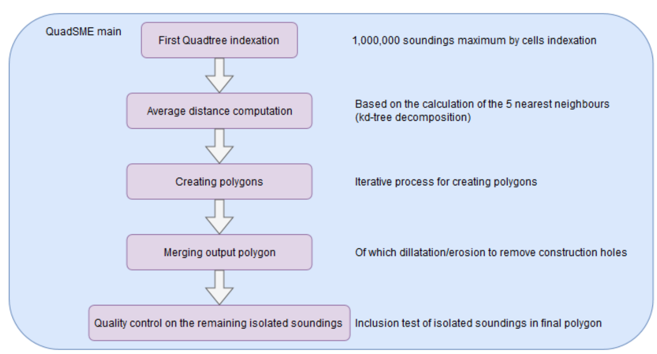

Fig. 2. Geographic extent check workflow.

Figures 3 shows the results obtained on a real dataset and in particular the integration of the complex geometry and holes that may exist in this bathymetric data coverage.

Preserving holes is essential when dealing with the selection of soundings originating from overlapping surveys, in order to prevent from an excessive filtering of valuable informa-

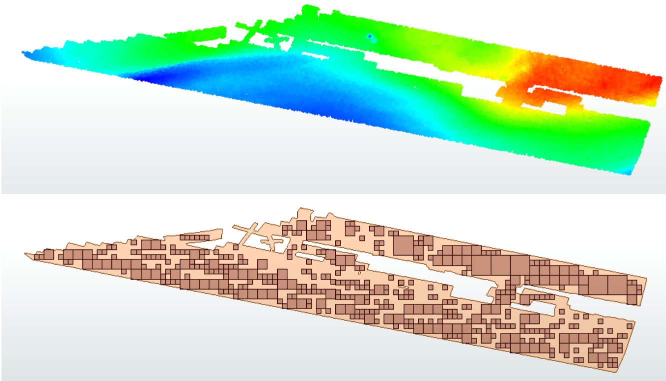

Fig. 3. On top a bathymetric survey (point cloud), on bottom its corresponding convex envelope with holes preserved. Also note the corresponding quadtree decomposition.

tion.This is why the SME must be as reliable as possible as shown in Figure 4

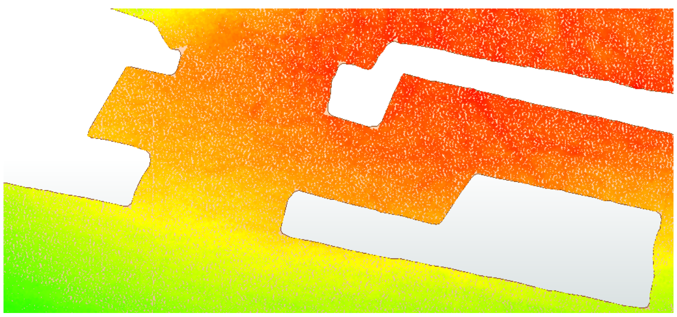

Fig. 4. Zoom on particular complex region of bathymetric survey.

D. Data clipping

Once the metadata and geometry of datasets are properly checked, the principle is to apply rules of deconfliction, defined in a matrix based on combinations of attributes (established by Shom hydrographic expert rules). The result may be to cut one or more previous surveys using SME as a cookie-cutter or not to cut (ie. the survey is added to complement an other survey in the area). This deconfliction process is performed on a one-by-one polygon.

Once all the polygons have been filtered, these polygons can be used to cut out the soundings from their surveys and integrate them into a spatial database (with a new associated SME) in order to be used for any type of bathymetric product.

E. Data quality control

Finally, a quality control process is carried out. The idea is to check that all relevant soundings of the nautical chart are included in the final bathymetric knowledge, or if not, can be explained: cartographic sounding selection, sounding not found in the bathymetric heritage... This inverse process is based on a tool to compare and search the equivalent soundings between the bathymetric knowledge and the cartographic knowledge (as the current nautical chart). It ensures a non-regression test. This quality control process has helped to improve the quality of the data, for example by detecting digitization errors. This control also allows for the reporting of Dangers to Navigation if critical information are not taken into account in the current navigation products.

IV. The first tile processing

The first deconfliction performed is shown in Figure 5. The area covered extends from the port of Saint-Malo in the west to the bay of Mont Saint-Michel in the east, and from the south of the river Rance to the town of Coutance in the north. The extent of this result (a tile of 1° by 1°) deals with 310 million soundings at the end, including more than 6 000 soundings digitized from old nautical charts (left). The result of the aforementioned automation process has raised 44 167 conflicts between overlapping data sources (right). Validation is undertaken by comparing, amongst other, with previously generated navigational product such as official ENCs (Electronic Navigation Chart).

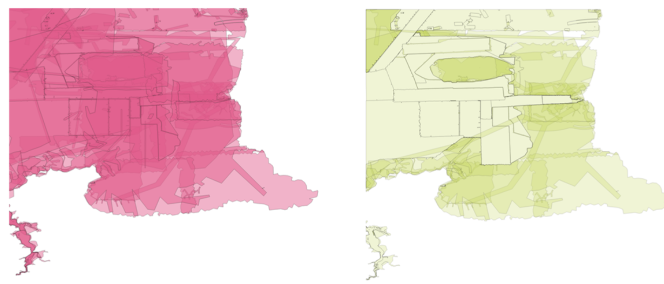

Fig. 5. First tile from the Téthys project, on the left the set of data studied for clipping and deconfliction, on the right the survey cut and keep after the deconfliction process

On this first tile (called 145_81, numbering inspired from the Marsden grid), it is 115 surveys that have been counted as input and 441,418,088 soundings to process. At the end of the all workflow (see Section III), 96 surveys were finally retained and 310,970,981 soundings integrated into the Téthys spatial database.

This first realization has recently benefited the cartographers when publishing a nautical chart covering Chausey islands and the production of a topo-bathymetric DTM covering the approaches of Saint-Malo, see [8].

V. Consultation portal and hydrographic process support tools

A. Overview

The Téthys data portal is an internal web GIS-based interface which provides fast and easy access to the bathymetric reference surface and associated information. The web-based user interface eliminates the need for a desktop GIS to view the data and thus improves the data accessibility.

Three components of the bathymetric reference surface are available for consultation (Figure 6):

- The spatial coverage of each survey present in the Téthys, represented as a vector polygon layer;
- The soundings, which are represented with a configurable color palette. Fluidity display is ensured by on the fly data decimation according to the zoom level.
- The metadata of selected bathymetric surveys (originator, positional and sounding accuracy, technique of sounding measurement, access restrictions …) are available through the attribute panel. Links to associated resources (e.g. hydrographic survey reports, plotting sheets) are provided for additional information.

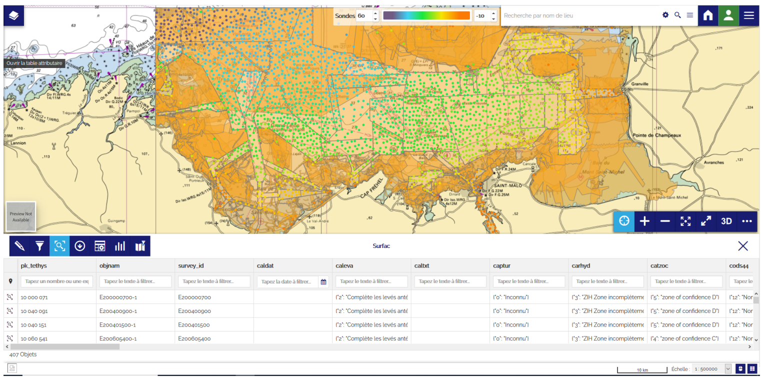

Fig. 6. Main page of Téthys data portal.

Selected layers can be queried, filtered and downloaded in well known GIS formats. Bathymetric data are extracted by defining a bounding box of the area of interest. Users can also load web services (WMS, WFS, WCS) or GIS vector data into the portal.

B. Spatial Data Infrastructure

Open source database management systems (DBMS) and web GIS packages are used to implement both the data management and visualization infrastructures. Figure 7 schematizes the architecture created, consisting of a PostgreSQL (13.2) database with PostGIS (3.1) and pdal (2.2.0) extensions, Geoserver (2.19) as map server and OpenLayers available in Mapstore 2 as web viewer.

A part of the point cloud data structure is based on pgPointcloud library, see [9]. This technology originally developed to manage high amount of 3D topographic LiDAR data has been adapted in Téthys workflow for bathymetric data, whether they are collected by acoustic sensors, LiDAR or mechanical means. Point cloud patch structuring offered by pgpointcloud library is a strong advantage to optimize the volume of this data and the querying.

C. Hydrographic survey planning support

The portal will allow to cross-referenced information from the Programme National d'Hydrographie (PNH, i.e. French National Hydrographic Plan), see [10], with current bathymetric knowledge in French waters. The PNH is the strategic document that defines the target of bathymetric knowledge at 4 years, according to various criteria, such as traffic, risks, recommended routes, environment... Direct access to knowledge makes it possible to monitor more accurately and in greater detail the progress of the implementation of the PNH. Thus, this monitoring guarantees a better visibility and prioritization of the new areas to be surveyed, for people in charge of survey planning.

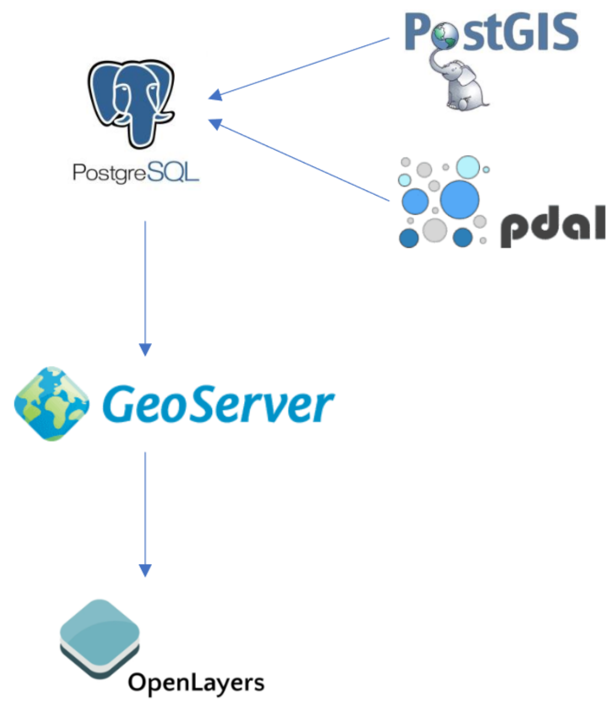

Fig. 7. Representation of the spatial data infrastructure.

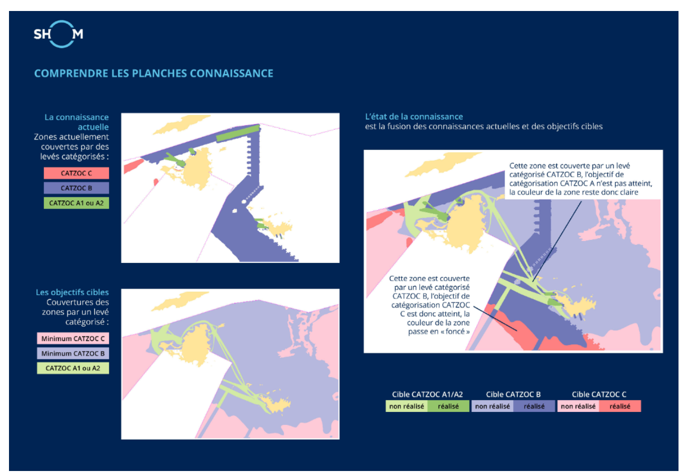

Fig. 8. National Hydrographic Plan [10]

D. Future Improvements

The Téthys data portal will give wide functionalities to users for generating DTMs from selected bathymetric data:

- interpolation of individual or aggregated bathymetric survey;
- the ability to apply a "background fill" option which populate data from global bathymetric models (GEBCO, EMOdnet ...) into any cells of the DTM that do not contain depth value;
- an option to choose grid resolution, interpolation method and associated parameters;
- an option to download the DTM in different formats (NetCDF, ESRI ascii grid and bag).

Each integration of a newly acquired bathymetric survey conducts to an automatic update of the bathymetric reference surface. In case of an update of the bathymetric layer in their extraction area, users will be notified by email. The subscription system will be directly configurable by each user.

VI. Conclusion and perspectives

The project is part of the foundation of the new bathymetric reference framework of Shom. It will prepare Shom to meet the challenges of the future: by better managing bathymetric data and making it more efficiently usable for end-products, the building of this reference framework strengthens the bathymetric processes in light of the significant increase in data foreseen as part of the undergoing program CHOF of the French hydrographic Service to renew its sounders and new techniques of hydrographic data acquisition and processing.

A lot of work remains to be done to optimize the management and the flows of this important mass of data in order to make this "Téthys" database completely interoperable with the Shom's hydrographic tools.

Moreover, in order to ensure an important control, many reliability actions are actually carried out manually but could be automated in the future in order to gain efficiency on the whole workflow, which would allow the Shom agents to concentrate on more complex problems.

Once the bathymetric reference built, a spatial coherence analysis of the data would identify soundings doubtful to be investigated or surveyed again. These coherence analysis tools, based on methods to be defined, have yet to be developed and are being studied and researched. This value-added capacity will enable to achieve a high level of qualification and meet its objectives as a reference framework.

The long-term objective is to put in place real governance of bathymetric data to ensure complete data transparency throughout its life cycle and to improve the reliability of use and the control of Shom processes in their completeness.

References

[1] IMO, "International convention for the safety of life at sea," tech. rep., International Maritime Organization (IMO), 1974. Available online: https://www.refworld.org/docid/46920bf32.html.

[2] P. A. Elmore and C. A. Steed, "Algorithm design study for bathymetry fusion- review of current state-of-the-art and recommended design approach," tech. rep., Naval Research Laboratory, December 2008.

[3] B. Calder, "On the uncertainty of archive hydrographic data sets," *IEEE Journal of Oceanic Engineering*, vol. 31, no. 2, pp. 249–265, 2006.

[4] A. Stateczny and I. Bodus-Olkowska, "Sensor data fusion techniques for environment modelling," in *2015 16th International Radar Symposium (IRS)*, pp. 1123–1128, 2015.

[5] L. Gomes and P. Oliveira, "Bathymetric data fusion: Pca based interpolation and regularization, sea tests, and implementation," in *OCEANS 2008*, pp. 1–8, 2008.

[6] S. M. Smith, "The navigation surface a multipurpose bathymetric database," Master's thesis, University of New Hampshire, 2003.

[7] R. Créach, S. Bosh, L. Boutry, G. Genevier Jonathan, P. Claverie, and B. Alexandre, "Scanbathy : A new solution to digitize depth data from historic survey sheets," in *11th Annual GEBCO Bathymetric Science Day (poster presentation)*, (Valparaiso, Chile), 12 October 2016.

[8] L. Biscara, C. Salvaterra, J. Rosetto, S. Le Jeune, H. Michaud, L. Seyfried, A. Pasquet, and C. Vrignaud, "Production de MNT topo-bathymétriques pour l'amélioration de la gestion du risque de submersion marine : cas du PAPI Saint-Malo," in *XVI Journées Nationales Génie Côtier*, pp. 727–734, 01 2020.

[9] P. Ramsey, P. Blottiere, M. Brédiff, and E. Lemoine, "pgPointcloud." https://github.com/pgpointcloud/pointcloud, 2013.

[10] Shom, "Programme national d'hydrographie 2021-2024," tech. rep., Shom, 2021. Available online: https://www.shom.fr/sites/default/files/2020-11/PNH_2021-2024_Web.pdf.