# Modèle de qualification interactif de données de commerce maritime imparfaites sur le XVIIIème siècle.

Directeurs de thèse : **Dr. HDR A Bouju, Dr C. Plumejeaud-Perreau**
Laboratoire/unité d'accueil : **UMR 7266 LIENSs, Université de la Rochelle; CNRS**
Ecole doctorale de rattachement : **ED 618 EUCLIDE**
Date limite de candidature : **30/04/2019**
Début de thèse : **1er octobre 2019**
Offre en ligne : https://emploi.cnrs.fr/Offres/Doctorant/UMR7266-CHRPLU-003/Default.aspx
Candidature : Cv, lettre de motivation et relevés de notes (M1&2) et rapports de stage, articles publiés le cas échéant à envoyer à christine.plumejeaud-perreau@univ-lr.fr , alain.bouju@univ-lr.fr

## Sujet de thèse

Ce sujet s'inscrit dans le cadre d'un programme financé par l'Agence Nationale de la Recherche, dénommé PORTIC, qui entend étudier les dynamiques spatiales et économiques à l'œuvre dans le processus de construction de marchés de plus en plus intégrés qui prépare et accompagne la Révolution industrielle. A cette fin, il croisera les données sur la navigation des ports français et celles issues de la balance du commerce afin de mieux saisir l'articulation entre espaces régionaux, nationaux et internationaux du commerce français du XVIIIe siècle, en s'appuyant sur deux corpus existants - Navigocorpus et Toflit18 – produits au cours de deux programmes ANR achevés. Le croisement des deux corpus permettra, entre autres, d'estimer plus précisément la part respective du commerce national et étranger, d'affiner les connaissances sur les ports qui articulent les marchés et leurs interactions, d'analyser les phénomènes régionaux de spécialisation entre plusieurs ports, de mesurer l'impact des conflits sur l'économie d'un port, de prendre la mesure de la contrebande à travers la Manche, de peser la part prise par les Français dans les services de transport international qui échappe aux statistiques commerciales de l'époque, ou encore de calculer la ratio entre la valeur du commerce et le tonnage ou les effectifs de main-d'œuvre affectés au transport maritime en fonction des flux.

PORTIC est un projet co-construit par des historiens, des économistes, des géomaticiens, des informaticiens, et des spécialistes de la communication de l'information par le Web, et qui vise à offrir des outils permettant une visualisation claire, scientifiquement irréprochable et calibrée pour des publics différents, d'informations historiques, en prenant pleinement en compte leur caractère imparfait. L'imperfection des données historiques dérive de lacunes documentaires, d'informations contradictoires délivrées par des sources différentes, ou de leur contenu imprécis. Ce caractère incertain d'une partie des informations, fondamental du point de vue de la compréhension du passé, est actuellement insuffisamment intégré par les outils de visualisation des données, notamment des flux. Les humanités numériques accompagnent toutes les étapes du projet, en permettant tout d'abord la mise en évidence des caractères aberrants et contradictoires des données par des outils de fouille et la mise en place de procédures interactives semi-automatisées par lesquelles les chercheurs qualifient la valeur des informations. Tout ce qui sera développé par PORTIC sera sous licence libre.

Ce projet de thèse aborde la question de la qualification de ces données avec une approche combinant à la fois des méthodes symboliques et numériques à travers un processus itératif intégrant les retours d'experts pour la curation des données du corpus.
Différents aspects seront abordés au cours de ce projet de thèse:
- Un modèle sémantique de trajectoires dérivé d'un modèle spatio-temporel générique (Tran et al. 2016) sera utilisé pour déduire des incohérences dans la base de données (informations contradictoires, itinéraires incohérents).
- Ce modèle sera connecté à un moteur exécutant des méthodes de fouille de données statistiques non paramétriques et non supervisées pour la détection de patrons récurrents et de valeurs aberrantes.
- Un modèle de qualité sémantique étendra le modèle sémantique actuel pour les trajectoires afin de gérer les annotations qualitatives.
- Les résultats seront affichés dans les interfaces de géo- visualisation de données (développées ailleurs dans le projet), permettant ainsi aux commentaires de l'expert d'être intégrés dans le modèle sémantique

pour une exploration itérative de différentes hypothèses. Cela implique un support pour un raisonnement non monotone en logique formelle de premier ordre.
L'approche sera évaluée tout d'abord en comparant d'anciens ensembles de données brutes avec les mêmes déjà corrigés manuellement, puis avec les données nouvellement collectées dans le projet en faisant en sorte que le logiciel interagisse avec les historiens possédant le rôle d'expert.

## Contexte de travail

Les fonds de la subvention seront gérés par le **CNRS**, qui sera l'employeur de l'étudiant. Le CNRS, dont la devise est «repousser la frontière de la science», couvre tous les domaines scientifiques, de l'étude de la matière et du monde vivant à celle des sociétés humaines. **L'école doctorale EUCLIDE** suivra le déroulement de la thèse.
Le doctorant sera hébergé dans les locaux dans l'Unité Mixte de Recherche **Littoral Environnement et Sociétés**, (U.M.R. 7266 LIENSs). Ce laboratoire regroupe des experts scientifiques des disciplines de l'écologie, la géographie, la biologie, l'histoire, la chimie moléculaire et les sciences de la terre et interroge des questions liées au développement durable et au changement climatique autour des zones littorales (https://lienss.univ-larochelle.fr/). Le doctorant intègrera donc un milieu fortement interdisciplinaire et en particulier le service de la plateforme base de données au croisement de nombreux projets scientifiques afin d'offrir une meilleure capacité de croisement de données fortement hétérogènes, et de favoriser la mise en œuvre des principes FAIR dans la recherche.
L'équipe de PORTIC sur la Rochelle est coordonnée par Christine Plumejeaud-Perreau, qui travaille depuis 5 ans avec Alain Bouju, Maitre de conférences avec Habilitation à Diriger des Recherches en informatique au **Laboratoire d'Informatique, Images et Interactions** (L3i) de l'Université de la Rochelle depuis 2014.
**Directeurs de thèse** : Alain Bouju (L3I) and Christine Plumejaud-Perreau (LIENSs).

- A. Bouju est un expert reconnu dans le domine de travaux sémantiques concernant les objets mobiles et l'étude de leur trajectoires, en particulier en ce qui concerne les mammifères marins et leur suivi par télémétrie (Wannous et al., 2017), ou le suivi de navires en mer (Etienne et al., 2009; Ray et al., 2015). Le L3i est reconnu internationalement dans le domaine des humanités numériques, et A. Bouju a ainsi participé à un projet ANR Alpage afin de proposer des modèles de métadonnées et de services adaptés à l'intégration et la diffusion de cartes anciennes imparfaites (Grosso et al., 2009).
- C. Plumejeaud a proposé durant son doctorat des méthodes de recherche de valeurs exceptionnelles interactives dans le domaine des statistiques socio-économiques (Plumejeaud et al., 2012), et elle a souvent été à l'interface d'historiens et d'informaticiens, notamment dans le cadre de son post-doctorat sur l'ANR GeoPeuple pour lequel elle a produit des interfaces Web interactives de géovisualisation permettant de comprendre l'histoire des communes et leur évolution démographique (Plumejeaud et al., 2014, Plumejeaud et al., 2015). Elle mène actuellement des recherches sur la qualification de données imparfaites, et donne des cours pour la communauté académique sur ce sujet lors d'Actions Nationales de Formation (Plumejeaud, 2018)

Le projet ANR PORTIC est une occasion pour le doctorant de se familiariser à **l'interdisciplinarité** dans **le domaine des humanités numériques** avec une équipe d'historiens de rang international, et des experts européens convoqués comme conseillés extérieurs au projet.
Par ailleurs, le modèle de curation interactif de PORTIC couplé au système de géovisualisation interactif de données imparfaites est réutilisable pour beaucoup de domaines et projets du laboratoire LIENSs. Ainsi, LIENSs est également un acteur du projet *Atlas historique de la Nouvelle-Aquitaine* qui propose entre autres un cas d'étude sur l'Amirauté de Marennes, situé en bordure de l'estuaire de la Seudre, très lié à l'économie maritime de la Rochelle, dans un milieu qui a énormément changé entre le 15 et le 16ème siècle. LIENSs est également un partenaire clé du projet régional DYPOMAR, pour l'étude multi-scalaire de la dynamique portuaire permettant de comprendre les transformations actuelles du littoral anthropisé.

## Contraintes et risques

Le fait que le corpus de données à analyser soit déjà présent et structuré dès le départ constitue une garantie pour le doctorant.
La collaboration étroite entre le L3i et le LIENSs (situés à 200 m l'un de l'autre) est également une garantie pour un dialogue fertile et, *in fine*, la bonne gestion de la thèse. En effet, Alain Bouju et Plumejaud-Perreau ont déjà co-dirigé un doctorant utilisant des technologies sémantiques (Tran, 2017) pour proposer une ontologie spatio-temporelle ayant des capacités de raisonnement et d'inférence logique, afin d'offrir un cadre générique pour une analyse et une visualisation croisées de données à long terme sur la biodiversité et les cultures.
Concernant l'avenir académique du doctorant, bien que le sujet soit réalisé dans un milieu interdisciplinaire, la thèse sera clairement soutenue en informatique (section 6 du CNRS, 27 du CNU, domaine de qualification des deux encadrants) et nous visons des publications dans des revues de rang international. Le doctorant disposera également de fonds pour rencontrer sa communauté de recherche, au niveau national comme international.

## Profil de candidature souhaitée

Formation : Master 2 spécialité Informatique / Ingénierie des connaissances
Expérience souhaitée en fouille de données (détections de similarités), Statistiques, Web sémantique et données liées (LOD).

## Bibliographie


Berti-Équille L., 2007, Quality Awareness for Managing and Mining Data, Habilitation à Diriger des Recherhes University of Rennes 1, France, http://pageperso.lif.univ-mrs.fr/~laure.berti/pub/Habilitation-Laure-Berti-Equille.pdf

Moreau, C., Devogele T. and Etienne L., 2018 Extraction de motifs de trajectoires sémantiques similaires. In : *Proceedings of Spatial Analysis and GEOmatics (SAGEO'2018)*, Eds : Mathieu Roche, Maguelonne Teisseire, Montpellier, 6-9 nov. 2018, France

Etienne L., Devogele T., Bouju A., 2009. « Analyse de similarité de trajectoires d'objets mobiles suivant le même itinéraire : Application aux trajectoires de navires », *Revue Ingénierie des Systèmes d'Information* (ISI), Hermès, vol. 14:5, p. 85-106.

Grosso E., Bouju A., Mustière S., 2009. "Data Integration GeoService: A First Proposed Approach Using Historical Geographic Data" In: *Proceedings of 9th International Workshop on Web and Wireless Geographical Information Systems*, Eds: J.D. Carswell, A. S. Fotheringham & G.McArdle, W2GIS 2009 7-8 December, Maynooth, Ireland Lecture Notes in Computer Science, pp 103-119.

Plumejeaud C., Villanova-Oliver M., 2012. "QualESTIM: Interactive Quality Assessment of Socioeconomic Data Using Outlier Detection", in: Gensel J., Josselin D., Vandenbroucke D. (Eds.), *Bridging the Geographic Information Sciences: International AGILE'2012 Conference*, Avignon (France), April, 24-27, 2012., Heidelberg, pp. 143–160. https://doi.org/10.1007/978-3-642-29063-3_8

Plumejeaud C., Grosso E., Parent B., 2014. 'Dissemination and geovisualisation of territorial entities' history', *Journal of Spatial Information Science*. doi:10.5311/JOSIS.2014.8.119 ; http://josis.org/index.php/josis/article/view/119

Plumejeaud C., Cristofoli P., Motte C., 2015. "De l'étude des nomenclatures territoriales à la modélisation des dynamiques des territoires administratifs en France", *Revue internationale de géomatique* 25, p. 355–392. https://doi.org/10.3166/rig.25.355-392

Plumejeaud-Perreau, C., 2018. La qualité des donnéées. In : Action Nationale de Formation '*Sciences des données*' [en ligne]. Sète. 6 novembre 2018. [Consulté le 29 novembre 2018]. Disponible à l'adresse : http://rbdd.cnrs.fr/IMG/pdf/qualite_des_donnees_plumejeaud_2018_04112018.pdf?517/365a13edab604bd0700b045bfac29a3607acb649

Ray C., Napoli A., Bouju A., Martin P.-Y., 2015. Detection of faked AIS messages and Resulting Risks, in: *IF&GIS 2015 - 7th International Workshop on Information Fusion and Geographic Information Systems*. Grenoble, France.

Tran B.-H., Bouju A., Plumejeaud-Perreau C., Bretagnolle V., 2016. "Towards a semantic framework for exploiting heterogeneous environmental data". *International Journal of Metadata, Semantics and Ontologies*, vol. 11, no 3, p. 191-205, https://doi.org/10.1504/IJMSO.2016.081586

Tran B.-H., 2017. *Une approche sémantique pour l'exploitation de données environnementales - Application aux données d'un observatoire*. Thèse de doctorat, Université de la Rochelle.
Wannous R., Malki J., Bouju A., Vincent C., 2017. "Trajectory ontology inference considering domain and temporal dimensions—Application to marine mammals". *Future Generation Computer Systems* 68, 491–499. https://doi.org/10.1016/j.future.2016.01.012

# Interactive quality assessment model for uncertain data about maritime trade during XVIIIth century

PhD supervisors: **Dr. HDR A Bouju, Dr C. Plumejeaud-Perreau**
Laboratory / reception unit: **UMR 7266 LIENSs, L3I, Université de la Rochelle; CNRS**
Graduate School: **ED EUCLIDE**
Deadline for application: **30/04/2019 – Starting :** Début de thèse : **1er october 2019**
Online offer : https://emploi.cnrs.fr/Offres/Doctorant/UMR7266-CHRPLU-003/Default.aspx
Application: CV, cover letter, certificate of achievement/ master's degree, notes and internship reports to send to christine.plumejeaud-perreau@univ-lr.fr, alain.bouju@univ-lr.fr

## Thesis subject

This subject takes place inside a program named PORTIC financed by French National Agency for Research (ANR). PORTIC intends to study the spatial and economic dynamics at work which resulted in increasingly integrated markets, a process which paved the way to, and sustained the Industrial Revolution. By crossing data on the shipping activities of French ports and those of the 18th-century French balance of trade, PORTIC aims at understanding the articulation between regional, national and international trade. PORTIC builds on two massive corpora produced by two previously achieved ANR programs, Navigocorpus and Toflit18. The combination of information contained in the two corpora will make it possible to estimate the respective part of national and international trade, to better perceive the ports which articulated markets and their interrelations, to analyze regional specialization mechanisms at work among different ports, to assess the importance of warfare on a port economy, to gauge the relevance of smuggling across the Channel, to quantify the part of French international transport services which is absent from contemporary statistics, to compute the ratio between trade value and tonnages or crew size in maritime transport flows.

PORTIC, a project jointly conceived by historians, economists, GIS specialists, computer scientists, and specialists in information and interaction design on the Web, will develop tools for a clear and scientifically sound visualization of these historical data, and their diverse reliability degree. Data's imperfection derives from missing information, contradictory data provided by different historical sources, and their imprecise content. The imperfect character of part of the information constitutes an essential element for a correct historical understanding, but it is presently insufficiently taken into account. Digital Humanities are fundamental to all the phases of the project. Data mining and semi-automatized procedures will make it possible to detect outliers and contradictory information and to let historians qualify its reliability.

This PhD project aims to tackle the qualification issues of those data with an approach combining both symbolic and numeric methods through an iterative process integrating experts' feedbacks for the corpus' data curation.
Different aspects will be dealt with during this thesis project:

- A semantic model for trajectories derived from a generic spatiotemporal model (Tran et al. 2016) will be used to deduce inconsistencies in the database (contradictory information, incoherent itineraries).
- This model will be connected with an engine executing some non-parametric and non-supervised statistical methods for patterns and outliers detection (data-mining field).
- A semantic quality model will extend the current existing semantic model for trajectories in order to handle the qualitative annotations.
- The findings will be displayed in the geo- and data-visualization interfaces (developed elsewhere inside the project), thus allowing for the expert's feedbacks to be integrated into the

semantic model for a further iterative exploration of various hypotheses. This will induce a support for non-monotonous reasoning.

The approach will be assessed firstly by comparing ancient raw datasets with the same already manually corrected, and then on newly collected data in the project by making the software interact with historians having the human expertise role.

## Research environment

Funds of the grant will be managed by the French National Center for Scientific Research (CNRS), who will be the employer of the student. CNRS whose motto is « pushing back the frontier of science », and covers all scientific fields from the study of matter and the living world to that of human societies. The EUCLIDE doctoral school will follow the management of the thesis.

The PhD will be based at the **Center for the Littoral, Environment, and Societies** (**LIENSs, Lab UMR 7266)**. LIENSs gathers expertise in different scientific fields (environmental, engineering, social sciences -geography and history- and humanities) to address issues related to the sustainable management of the coastal zone (https://lienss.univ-larochelle.fr/).

PORTIC's interactive data curation model coupled with the geo-visualization framework for imperfect data is valuable for many of the fields and projects in this research center. For instance, LIENSs is part of the regional project Atlas historique de la Nouvelle-Aquitaine, which features a case study focusing on the Marennes district: a small administrative capital (admiralty, election, bailiwick and then arrondissement) located on the estuary of the Seudre, which was linked to La Rochelle's maritime economy, in an environment that underwent major changes since the 15th-16th centuries. LIENSs is also a key partner of the DYPOMAR project, as the multi-scalar study of port dynamics is essential to understand present-day coastal transformations.

The PORTIC LIENSs team is coordinated by Christine Plumejeaud-Perreau, who has been working for the DYPOMAR project with Alain Bouju of the **Laboratory for Computing, Images, and Interaction (L3i)** at University of La Rochelle since 2014. This team will be in charge of the implementation of a progressive, user-friendly interface for the geo-visualization of maritime flows, and of the integration of a semantic data model dedicated to trajectories for detecting logical inconsistencies. The close collaboration that already exists in La Rochelle between LIENSs and L3I (200 meters away from) is a guarantee for a good management of the thesis.

**Supervisors :**

- A. Bouju (HDR) is an expert of semantic works on trajectories, processing for instance marine mammals' telemetric data (Wannous et al., 2017), ships' mobility (Etienne et al., 2009; Ray et al., 2015). His research centre, L3i, has an internationally recognized competence for handling cultural heritage in digital systems, and A. Bouju has worked on Alpage project for handling metadata and services for integration of old imperfect maps (Grosso et al., 2009).
- C. Plumejeaud has worked on interactive outliers' detection and long-term database building during her thesis (Plumejeaud et al., 2012), and she has often been the interface between historians and computer scientists, such for instance for the ANR GeoPeuple program, that produced geo-visualisation interfaces for understanding municipalities' history (Plumejeaud et al., 2014, Plumejeaud et al., 2015).

PORTIC's interactive data curation model coupled with the geo-visualization framework for imperfect data is valuable for many of the fields and projects in this research center. For instance, LIENSs is part of the regional project *Atlas historique de la Nouvelle-Aquitaine*, which features a case study focusing on the Marennes district: a small administrative capital (admiralty, election, bailiwick and then arrondissement) located on the estuary of the Seudre, which was linked to La Rochelle's maritime economy, in an environment that underwent major changes since the 15th-16th centuries. LIENSs is also a key partner of the DYPOMAR project, as the multi-scalar study of port dynamics is essential to understand present-day coastal transformations.

## Constraints and risks

The corpus of data to be analyzed is already present and structured and this is a guarantee for the doctoral student.

The close collaboration between the L3i and the LIENSs (located 200 m apart) is also a guarantee for a fertile dialogue and, ultimately, the good management of the thesis. Alain Bouju and Plumejaud-Perreau already co-supervised a PhD. thesis using semantic technologies (Tran, 2017), with a particular spatiotemporal ontology having effective capabilities for reasoning and logical inference, in order to offer a generic framework for a cross-analysis and visualization of long-term biodiversity and crop data.
Regarding the academic future of the doctoral student, although the subject is carried out in an interdisciplinary environment, the thesis will be clearly supported in computer science (section 6 of the CNRS, 27 of the CNU, qualification area of the two supervisors) and we aim publications in international journals. The doctoral student will also have funds to meet his research community, at both nationally and internationally levels.

## Candidate profile

Education: Master specialty Computer Sciences / Knowledge engineering
Desired Experience in Statistics and Data Mining (outliers detections, patterns recognition), Semantic Web and Linked Data.

## References


Berti-Équille L., 2007, Quality Awareness for Managing and Mining Data, Habilitation à Diriger des Recherhes University of Rennes 1, France, http://pageperso.lif.univ-mrs.fr/~laure.berti/pub/Habilitation-Laure-Berti-Equille.pdf

Moreau, C., Devogele T. and Etienne L., 2018 Extraction de motifs de trajectoires sémantiques similaires. In : *Proceedings of Spatial Analysis and GEOmatics (SAGEO'2018)*, Eds : Mathieu Roche, Maguelonne Teisseire, Montpellier, 6-9 nov. 2018, France

Etienne L., Devogele T., Bouju A., 2009. « Analyse de similarité de trajectoires d'objets mobiles suivant le même itinéraire : Application aux trajectoires de navires », *Revue Ingénierie des Systèmes d'Information* (ISI), Hermès, vol. 14:5, p. 85-106.

Grosso E., Bouju A., Mustière S., 2009. "Data Integration GeoService: A First Proposed Approach Using Historical Geographic Data" In: *Proceedings of 9th International Workshop on Web and Wireless Geographical Information Systems*, Eds: J.D. Carswell, A. S. Fotheringham & G.McArdle, W2GIS 2009 7-8 December, Maynooth, Ireland Lecture Notes in Computer Science, pp 103-119.

Plumejeaud C., Villanova-Oliver M., 2012. "QualESTIM: Interactive Quality Assessment of Socioeconomic Data Using Outlier Detection", in: Gensel J., Josselin D., Vandenbroucke D. (Eds.), *Bridging the Geographic Information Sciences: International AGILE'2012 Conference*, Avignon (France), April, 24-27, 2012., Heidelberg, pp. 143–160. https://doi.org/10.1007/978-3-642-29063-3_8

Plumejeaud C., Grosso E., Parent B., 2014. 'Dissemination and geovisualisation of territorial entities' history', *Journal of Spatial Information Science*. doi:10.5311/JOSIS.2014.8.119 ; http://josis.org/index.php/josis/article/view/119

Plumejeaud C., Cristofoli P., Motte C., 2015. "De l'étude des nomenclatures territoriales à la modélisation des dynamiques des territoires administratifs en France", *Revue internationale de géomatique* 25, p. 355–392. https://doi.org/10.3166/rig.25.355-392

Plumejeaud-Perreau, C., 2018. La qualité des donnéées. In : Action Nationale de Formation '*Sciences des données*' [en ligne]. Sète. 6 novembre 2018. [Consulté le 29 novembre 2018]. Disponible à l'adresse : http://rbdd.cnrs.fr/IMG/pdf/qualite_des_donnees_plumejeaud_2018_04112018.pdf?517/365a13edab604bd0700b045bfac29a3607acb649

Ray C., Napoli A., Bouju A., Martin P.-Y., 2015. Detection of faked AIS messages and Resulting Risks, in: *IF&GIS 2015 - 7th International Workshop on Information Fusion and Geographic Information Systems*. Grenoble, France.

Tran B.-H., Bouju A., Plumejeaud-Perreau C., Bretagnolle V., 2016. "Towards a semantic framework for exploiting heterogeneous environmental data". *International Journal of Metadata, Semantics and Ontologies*, vol. 11, no 3, p. 191-205, https://doi.org/10.1504/IJMSO.2016.081586

Tran B.-H., 2017. *Une approche sémantique pour l'exploitation de données environnementales - Application aux données d'un observatoire*. Thèse de doctorat, Université de la Rochelle.

Wannous R., Malki J., Bouju A., Vincent C., 2017. "Trajectory ontology inference considering domain and temporal dimensions—Application to marine mammals". *Future Generation Computer Systems* 68, 491–499. https://doi.org/10.1016/j.future.2016.01.012