h e g

Haute école de gestion
Genève

Évaluation de l'impact de l'utilisation de méthodes de fouille de données pour améliorer la qualité de l'information du trafic maritime de matière première

Travail de master réalisé dans le cadre du master en Sciences de l'information

Par :

David Dällenbach

Sous la direction de :

Arnaud GAUDINAT, Professeur HES

Genève, le 17 août 2020

Haute École de Gestion de Genève (HEG-GE)

Filière Sciences de l'information

Hes·so GENÈVE
Haute Ecole Spécialisée de Suisse occidentale

Déclaration

Ce travail de Master est réalisé dans le cadre de l'examen final de la Haute école de gestion de Genève, en vue de l'obtention du titre de Master of Science en Sciences de l'information.

L'étudiant a envoyé ce document par email à l'adresse remise par son directeur de travail de Master pour analyse par le logiciel de détection de plagiat URKUND, selon la procédure détaillée à l'URL suivante : http://www.urkund.fr/student_gorsahar.asp

L'étudiant accepte, le cas échéant, la clause de confidentialité. L'utilisation des conclusions et recommandations formulées dans le travail de Master, sans préjuger de leur valeur, n'engage ni la responsabilité de l'auteur, ni celle du directeur du travail de Master, du juré ou de la HEG.

« J'atteste avoir réalisé seul le présent travail, sans avoir utilisé des sources autres que celles citées dans la bibliographie. »

Fait à Gland, le 17.08.2020

David Dällenbach

Source de l'image de titre :

LÜCKMANN, D., 2020. Container terminal in Hambourg (non-commercial) [en ligne]. [Consulté le 1 août 2020]. Disponible à l'adresse : https://unsplash.com/photos/SInhLTQouEk

Remerciements

Accompli dans le contexte particulier de la pandémie du COVID-19, la réalisation de ce travail de mémoire a nécessité beaucoup d'énergie et de résilience ; l'impossibilité d'accéder à certaines ressources durant une bonne partie de sa conception et de sa rédaction, devoir lutter contre les tentations qui découlent du *Home Office* et les recommandations d'éviter les contacts avec les différentes personnes impliquées dans ce projet, font de ce mémoire un moment particulier dans ma vie. Je tiens tout particulièrement à remercier :

Arnaud Gaudinat, pour son soutien, sa disponibilité et ses conseils pendant les jours de tempêtes.

François Can, pour ce sujet passionnant et très stimulant.

Guy Druey, pour les moments d'échanges et les données sans lesquelles mon travail aurait été bien différent.

Mes collègues de travail qui ont souvent manifesté beaucoup de soutien et d'intérêt lors de ce long processus de mémoire.

Adrien Dubied, pour sa relecture attentive et son soutien.

Pablo Crotti, pour ses inputs et son amitié malgré la distance.

Mes amis, les Fous, pour tous les moments formidables vécus en leur présence, depuis plus de vingt ans.

Ma famille, pour l'indéfectible confiance qu'elle me témoigne depuis toutes ces années et, particulièrement ma femme, sans qui beaucoup de projets et de rêves n'auraient été réalisés. Mais aussi ma fille qui me permet d'être « le meilleur du monde entier ».

Et enfin mon père, qui ne connaîtra pas l'aboutissement de mon travail, mais qui a toujours compris mon besoin d'explorer le monde.

Sigles et abréviations

AIS	Automatic Identification Service
S-AIS	Satellite-AIS
DQ	Data Quality
DQA	Data Quality Assessment
DQD	Data Quality Dimension
ETA	Estimated Time of Arrival
ETD	Estimated Time of Departure
GPS	Global Positioning System
HDG	Heading
IMO	International Maritime Organization
ITU	International Telecommunication Union
MMSI	Maritime Mobile Service Identity
RL	Riverlake
ROT	Rate Of Turn
SARA	Sharing AIS-Related Anomalies
SOLAS	Safety of life at sea (Convention Internationale de 1974)
TMMP	Trafic maritime de matières premières
UN/LOCODE	United Nations Code for Trade and Transport Locations
VHF	Very High Frequency
VTS	Vessel traffic services
WM	Web Mining

Résumé

À l'ère où le volume, la variété et la vélocité des données sont en constante augmentation, le développement d'outils de scraping et l'exploitation des données disponibles dans le Web public représentent non seulement une opportunité formidable, mais également un défi de taille pour les entreprises qui souhaitent les exploiter. Alors que celles-ci sont utilisées dans la prise de décisions, la qualité de ces données est d'une importance cruciale. Une fois collectées, l'utilisateur doit s'assurer que ces données sont fiables et pertinentes. Une évaluation objective permet à l'utilisateur de cerner le potentiel des données tout en limitant les coûts quant à leur traitement ainsi que la prise de décisions reposant sur des données erronées.

Mandaté par Riverlake, ce travail a pour objectif de proposer une méthodologie d'évaluation de la qualité des données issues de onze ports commerciaux et d'en mesurer l'impact sur les données déjà à la disposition de l'entreprise grâce au système AIS. Dans sa première partie, cette étude passe en revue les spécificités et les limites de ce système. Ensuite, il est question d'appréhender ce que signifie la qualité des données et comment cette dernière peut être évaluée à l'aide d'un processus de *Data Quality Assessment* (DQA). Reposant sur les dimensions de la qualité identifiées dans la littérature scientifique, et confrontées aux besoins de notre mandat par le biais d'un questionnaire, sept dimensions ont finalement été mesurées dans le but de fournir une évaluation globale chiffrée des données récoltées sur le Web public. Avec un résultat de 79,74 %, cet indice global de qualité démontre que les données récoltées sont efficaces dans l'amélioration et l'enrichissement des données à disposition de Riverlake. Cependant, l'absence de benchmarks pour comparer les résultats de cette étude nuance sa portée.

Sur la base de cette étude et des limites décelées, nous recommandons de : réfléchir à la création des benchmarks tant pour évaluer les dimensions que pour l'indice globale, identifier des pondérations pour l'indice global qui soient en accord avec les besoins et le contexte de travail de notre mandant, considérer le traitement des données après scraping comme une étape importante de l'amélioration de la qualité générale des données et enfin que les sources utilisées lors d'un scraping doivent être évaluées afin d'en limiter le traitement.

Mots clefs : Big Data, Web Scraping, Data Quality, Data Quality Assessment, Data Quality Dimension, Data Quality Metrics, Weighted Metrics, Traffic Maritime de Matières Premières, Indice Global de Qualité, AIS, Precise Intelligence

Table des matières

Liste des tableaux

Liste des figures

1. Introduction

1.1 Contexte

Le trafic maritime de matières premières (TMMP) n'échappe pas au phénomène du Big Data et à l'utilisation d'une quantité de données de plus en plus importantes à des fins aussi diverses que variées, allant de la sécurité en mer à l'avantage concurrentiel pour les entreprises de négoce (Serry & Lévêque 2015, Yang 2019). Principalement issues du Automatic Identification System (AIS) qui permet l'échange automatisé de messages entre les navires ainsi que les différentes autorités par le biais de satellites et d'antennes relais, ces données s'avèrent cependant peu ou pas fiables pour toutes tentatives visant à les exploiter dans d'autres contextes et avec d'autres objectifs que la sécurité et la sûreté en mer (Adland 2017, Can, Gaudinat & Theodoro 2020, Harati-Mokhtari 2007).

Afin de parer au manque de confiance et de précision des données AIS, il est désormais possible de récolter, par le biais de différentes méthodes de Web Mining (WM), des données éparses issues de différentes sources internet dans le but de les compiler, de les exploiter et d'en extraire de la connaissance (Berti-Equille 2004, Tu 2017). En procédant de la sorte, les bases de données propriétaires peuvent être enrichies et améliorées afin de devenir de meilleurs instruments d'aide à la décision. Or, ces données sont très souvent proposées de manières hétérogènes, non structurées et redondantes, impactant non seulement leur qualité, mais surtout leur pertinence et, limitant donc le potentiel et les opportunités dont elles recèlent. À ce jour, aucune recherche n'a été menée sur la qualité des données récoltées sur les sites internet des ports commerciaux, dans le cadre du trafic maritime de matières premières. Il est donc impératif d'évaluer la qualité des données issues du WM et ainsi l'apport de ces dernières aux données issues du système AIS.

1.2 Mandat

C'est dans le contexte énoncé ci-dessus que s'inscrit le mandat qui nous a été confié par la société genevoise de logistique et de courtage maritime Riverlake Shipping SA[1] (RL). Par le biais du projet Innosuisse « *Precise Intelligence : Big Data Analytics for Comprehensive Global Trade Flow Intelligence »* (Can, Gaudinat et Theodoro 2020) qui associe la HES-SO Genève et RL, l'objectif est d'améliorer la qualité des données sur les flux commerciaux des marchandises avec des informations provenant de sources multiples, et plus particulièrement du web public, dans le but de développer la prévision du TMMP.

[1] https://www.riverlake.ch/en/riverlake-group.php

Les informations en possession de RL reposent à l'heure actuelle essentiellement sur les données provenant du système AIS ainsi que les données de navigation issues d'un prestataire externe. Bien que les données AIS aient démontré toute leur utilité dans le cadre de la sécurité et la sûreté en mer, certaines informations liées à ces dernières comme le temps estimé d'arrivée (ETA), le nom du port d'arrivée ou encore le type de matière première sont encore peu fiables, lacunaires, voire totalement inconnues. La mise à disposition de certaines données sur internet par des ports commerciaux ou des agences maritimes permet de combler certaines de ces insuffisances. Grâce à des outils de Web scraping capables d'extraire de l'information de ces sites, il est dorénavant possible d'améliorer la qualité des données à disposition de RL.

Ce travail vise donc à mener une réflexion de fond sur la qualité objective des données issues de méthode de fouille de données textuelles et leur apport sur la qualité globale des données à disposition de RL. Sa finalité réside d'une part dans l'identification de métriques adaptées à la conduite d'une évaluation de la qualité, d'effectivement procéder à cette évaluation pour mesurer l'impact de la fouille de données et d'autre part, de proposer une méthode d'évaluation contextuelles adaptée aux besoins et aux objectifs de l'entreprise.

Pour cela, ce dossier s'articule en deux parties : de par son caractère exploratoire, cette étude nécessite tout d'abord un état de l'art qui décrit les particularités du système AIS, ses apports et ses limites. Ensuite, cette contextualisation vise la mise en relief, le rassemblement et la discussion des concepts développés dans la littérature scientifique à propos de la qualité des données. Ceci non seulement dans le but d'appréhender comment évaluer la qualité, mais aussi de créer un modèle permettant de répondre à nos questions de recherches. Ces dernières sont justement traitées dans la deuxième partie de ce dossier. Grâce à un échantillon stratifié provenant de l'étude conduite simultanément à cette recherche (Druey 2020), nous testerons le modèle imaginé avant d'émettre des recommandations utiles à l'ajustement de notre champ d'étude et l'amélioration du projet.

1.3 But de la recherche

1.3.1 Objectifs

Ce travail de master vise les objectifs généraux suivants :

- Réaliser un état de l'art de l'évaluation de la qualité des données du TMMP.
- À partir de la littérature et du référentiel qui sera réalisé dans le second travail de master en lien avec ce mandat intitulé « *Etude, conception, collecte, curation et évaluation d'un scraping de sites web liés au transport maritime pour*

améliorer la prédiction du fret de matières premières » par Guy Druey, identifier les métriques pertinentes pour évaluer la qualité des données du TMMP.

- Mettre au point une méthodologie d'évaluation permettant de clarifier l'apport ou non de la fouille de données textuelles (Web mining et rapport par mail).
- Faire une évaluation sur un échantillon représentatif
- Faire des recommandations sur les méthodologies d'évaluation de la qualité de l'information du TMMP et sur l'utilisation des méthodes de fouilles de données dans ce contexte.

1.3.2 Question de recherche

De nos objectifs généraux découlent deux questions de recherche :

- Quelles sont les métriques pertinentes quant à l'évaluation de la qualité des données ?
- Est-ce que les données récoltées par le biais d'outils de Web scraping auprès des sites internet de certains ports permettent d'améliorer la qualité des données ?

2. Automatic Identification System

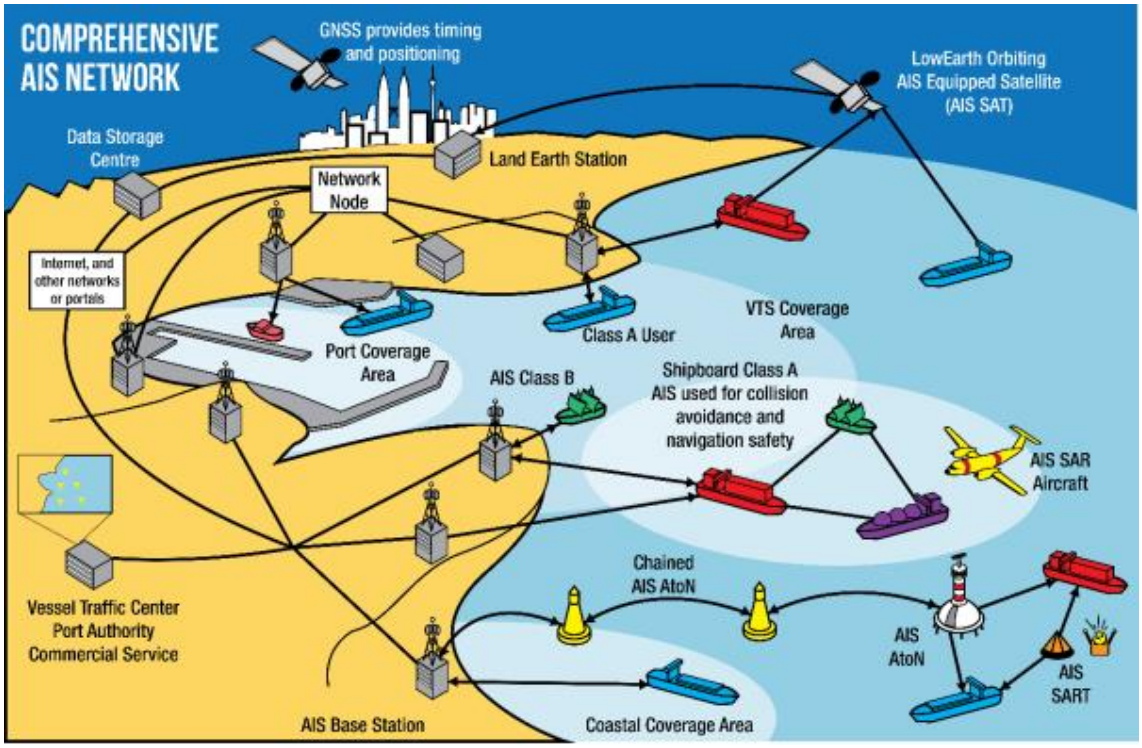
2.1 Présentation AIS

Dans un contexte d'intensification des échanges commerciaux, de transport ou des activités humaines liées à la pêche ou à la plaisance sur les océans, l'augmentation constante du trafic maritime dans toutes les régions du globe a nécessité la mise en place de nouveaux outils de surveillance. Ceux-ci permettent d'accroître non seulement la sécurité et la sûreté des équipages, mais aussi d'améliorer la navigation et sa gestion, tout en garantissant la protection de l'environnement contre d'éventuelles catastrophes (Serry & Lévêque 2015, p. 181). Auparavant, la navigation reposait principalement sur des technologies passives telles que le radar dont l'utilisation était parfois sensiblement altérée par de mauvaises conditions météorologiques, des capacités de détection et d'interprétation des dangers limitées ou des phénomènes naturels provoquant des anomalies (Felski & Jaskolski 2012, p.17). Il était donc nécessaire de développer de nouveaux systèmes tels que le Vessel Monitoring System (VMS), le Long Range Identification System (LRIT) ou l'Automatic Identification System (AIS) afin d'apporter plus de précision aux instruments existants et de contribuer plus efficacement non seulement à l'identification des navires, mais aussi à leur localisation afin de diminuer les risques de collision, d'accident ou de catastrophe écologique.

Dans ce contexte, l'*International Maritime Organization* (IMO) a décidé d'imposer progressivement l'utilisation du Système d'identification automatique ou *Automatic Identification System* (AIS) dès 2001. Cette volonté formalisée dans la convention *Safety*

of Life At Sea (SOLAS) de 2002 et l'installation de ce nouveau dispositif rendu obligatoire par l'IMO dès 2004 pour les navires de type classe A soit « *All ships of 300 gross tonnage and upwards engaged on international voyages and cargo ships of 500 gross tonnage and upwards not engaged on international voyages and passenger ships irrespective of size shall be fitted with an automatic identification system*[2] » (IMO, 2004). Basé sur l'échange automatisé de messages par radio *very high frequency* (VHF), l'objectif principal du système AIS est la maîtrise de la circulation maritime et sa sécurisation. Ainsi, en fournissant des informations (pratiquement) en temps réel de navire à navire ou de navire aux autorités de surveillance et de secours par le biais d'un réseau de communication constitué notamment de satellites (S-AIS[3]) et d'antennes relais, le trafic maritime est entré dans l'ère de la E-Navigation[4] ou « d'une situation maritime enrichie » (Serry & Lévêque 2015, p. 181).

Figure 1 : Le réseau AIS



(Ministry of Transport, Malaysia, s.d.)

[2] La classe B correspond aux petits navires non soumis aux conventions SOLAS (plaisance, navire de pêche de moins de 15 mètres).

[3] Jusqu'en 2009, la visibilité d'un navire par le biais du système AIS se limitait à la présence d'un récepteur dans un rayon de 20 milles marins (Serry & Lévêque 2015, p. 180).

[4] L'IMO définit l'E-Navigation comme la création, la collecte, l'intégration, l'échange et la présentation harmonisés d'informations maritimes à bord et à terre par voie électronique visant à améliorer la navigation, la sécurité, la sûreté en mer et la protection du milieu marin (IMO, 2008)

2.1.1 Fonctionnement du système AIS

Le système AIS est un transpondeur autonome installé sur les navires qui émet et reçoit en continu. Connecté aux autres instruments de navigation (GPS, compas gyroscopique, un ordinateur de bord, etc.), ce dernier est muni d'un écran de contrôle (Display Control Unit) permettant de configurer le système lors de son installation et d'interagir avec celui-ci par la suite.

Le système AIS permet l'échange d'informations relatives à la navigation et au voyage d'un navire émises à destination des autres navires, des stations côtières ou des autorités maritimes (Figure 1). Celui-ci peut gérer plusieurs rapports d'émission et de réception de données simultanément, à différents taux de mise à jour, qui varient en fonction d'une part, de la vitesse du navire et d'autre part, de son statut de navigation (Annexe 4). Les informations issues du système AIS se composent de différents types de données classées en 4 catégories distinctes, soit statiques, dynamiques, liées au voyage ou à la sécurité (Annexe 2). Les données dites statiques sont entrées dans le système lors de l'installation et ne doivent être modifiées que si le navire change par exemple de nom, d'indicatif d'appel ou de numéro d'identification comme le *Maritime Mobile Service Identity* (MMSI). En ce qui concerne les données dynamiques, ces dernières seront automatiquement mises à jour via les capteurs du navire directement connectés au système AIS (GPS, etc.), et enfin les données liées au voyage sont saisies manuellement lors de chaque voyage par le membre d'équipage assigné à cette tâche (Harati Mokhtari & al. 2007, p. 374).

Globalement, les informations récoltées par le système AIS sont transmises sous forme de messages dont les normes et la signification des valeurs dans les champs ont été définies par l'ITU (ITU, 2014) (Annexe 3-4). Il existe 27 différents messages regroupant différentes combinaisons de champs de données (statiques, dynamiques, etc.) en fonction du but de la transmission (Annexe 1). Parmi ces 27 types de messages, ceux en lien avec le report de positions sont bien évidemment les plus utilisés au vu de leur importance pour la navigation, mais aussi pour la richesse des données qu'il est possible d'en extraire a posteriori.

3. Revue de littérature

3.1 Données maritimes, un champ de recherche qui a le vent en poupe

Désormais, comme pratiquement dans tous les domaines de la vie et des activités humaines, le trafic maritime n'échappe pas à la vague du Big Data. En effet, dans leur

article paru en 2019, Yang & al. offrent une synthèse des recherches en matière de nouveaux services et d'applications de la technologie AIS liées à l'émergence de ce que certains experts nomment dorénavant la marétique[5], et soulignent l'augmentation des publications scientifiques traitant du sujet depuis une dizaine d'années (Yang & al. 2019, p. 766). À sa création, le système AIS a été conçu uniquement dans le but d'éviter les collisions entre navires et ainsi, d'améliorer la sécurité en mer. Mais les données disponibles durant les premières années de son fonctionnement se limitaient à la capacité des communications par ondes radio VHF d'une porte maximale de 10-20 nautiques par rapport aux côtes[6].

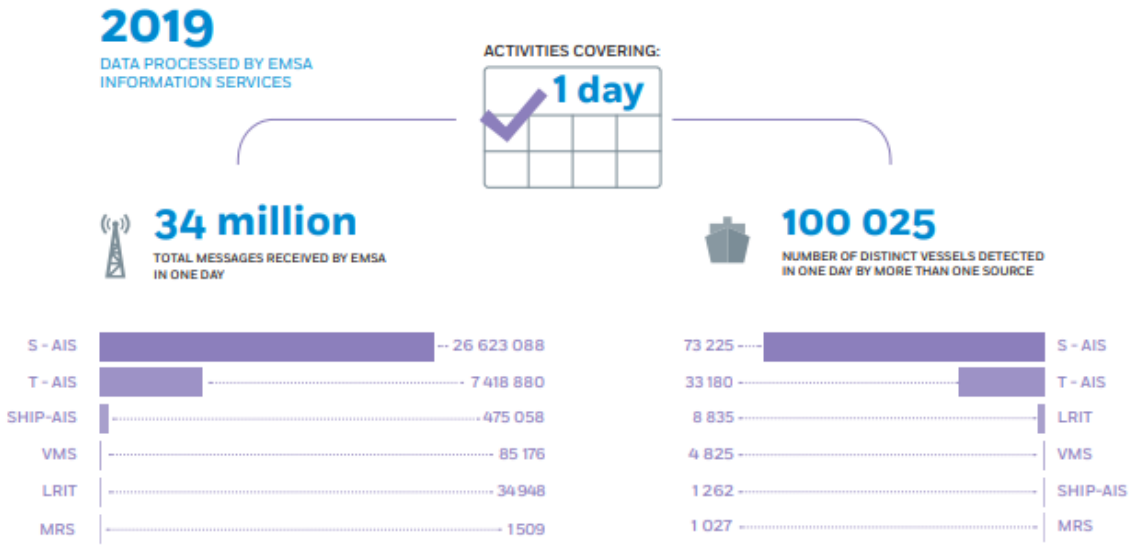
Or depuis 2008, le lancement de satellites équipés de récepteur AIS pouvant capter les données transmises par les dispositifs embarqués sur les navires, a radicalement changé la manière dont le système AIS fonctionne. Ces satellites permettent d'avoir une couverture beaucoup plus performante tout en réduisant sensiblement les zones blanches situées en haute mer. De plus, les récepteurs terrestres AIS, situés sur les côtes ainsi que dans les ports, étaient très souvent surchargés en raison du trop grand volume de données et donc dans l'impossibilité de remplir leur fonction. Le réseau satellitaire a donc augmenté les capacités du système AIS tout en offrant un moyen plus simple et plus rapide de collecter des données à l'échelle du globe, presque en temps réel. Dès lors, avec l'amélioration du système AIS et la diversité des informations qu'il contient et transmet, les applications possibles des données AIS ne se limitent plus uniquement à la sécurisation de la navigation et à la surveillance maritime, mais s'étendent à de nombreux autres champs d'utilisation (Yang & al. 2019, p. 756).

Les recherches exploratoires menées pour ce travail de master présentent la grande diversité des domaines d'étude impactés par l'opportunité qu'incarnent les systèmes AIS et son apport prometteur à certaines disciplines. En effet, l'utilisation de satellites a permis de passer d'une couverture limitée à la proximité des côtes à une couverture globale provoquant ainsi une explosion des données disponibles (Figure 2). Ainsi, il y a désormais un volume de données suffisant pour développer par exemple des algorithmes de machine learning pour analyser le trafic maritime et en dégager des modèles de navigation. (Cazzanti & Pallotta 2015, Pan & Deng, 2009, Tu & al. 2018, Shelmerdine 2015). Profitant de l'apport des données AIS satellitaires, nous pouvons également citer les travaux qui se sont penchés sur la prédiction de trajectoires et l'analyse des mouvements des navires (Mao & al. 2018, Redoutey & al. 2008, Last & al. 2014, Arguedas & al. 2018, Deng & al. 2014) ou l'estimation de modèles de navigation

[5] https://issuu.com/opteam/docs/seagital-livre-bleu-12112013
[6] Entre 20 à 40 kilomètres.

(Aarsæther & Moan 2009). Enfin, nous ne pourrions donner un rapide panorama des nouvelles applications du système AIS sans évoquer par exemple les études qui traitent des risques environnementaux (Kaluza & al. 2010, Winther & al, 2014) et de l'impact du trafic maritime sur l'écosystème marin (Gerritsen & al. 2013).

Figure 2 : Données traitées par l'EMSA par jour



(The EMSA Facts and Figures 2019, p.10)

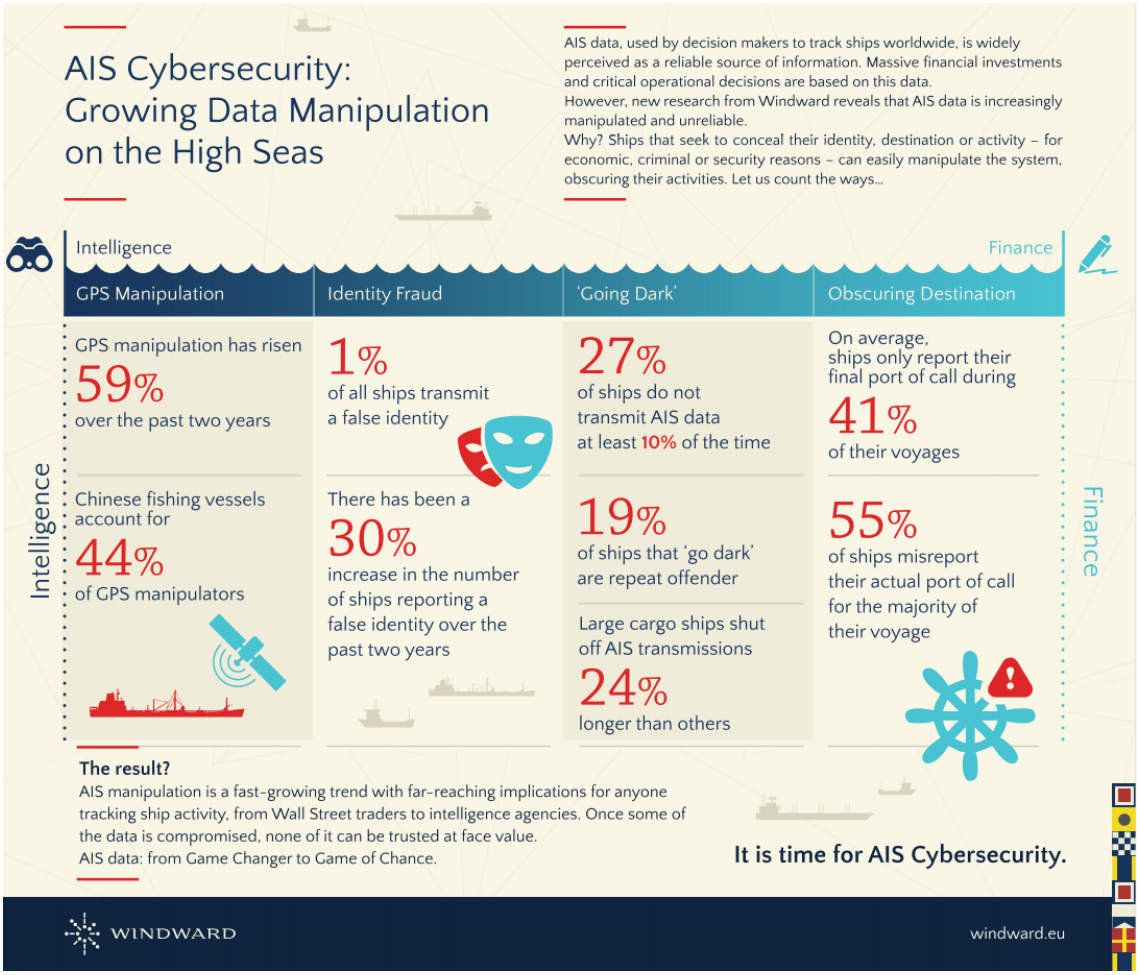
3.2 L'écueil de la qualité

Malgré des recherches qui démontrent sa contribution à la résolution de nouvelles problématiques, une part non négligeable de la littérature étudiée dans le cadre de cette recherche pointe les erreurs du système AIS (Annexes 6-10). Ces dernières peuvent s'appréhender par leur intentionnalité ou par leur non-intentionnalité (Iphar & al. 2015).

Dans le premier cas, des chercheurs se sont par exemple intéressés aux menaces extérieures qui pèsent sur la sécurité du système AIS et en particulier sur l'éventualité d'attaques informatiques sur les logiciels AIS ainsi que sur les fréquences radio réservées pour la transmission des messages. Consistant soit à émettre de fausses informations soit à complètement brouiller l'échange d'information (*spoofing*), ces activités ont pour conséquence d'induire les membres d'équipage en erreur dans le but par exemple de les attaquer (faux message de détresse, navire fantôme, etc.) (Balduzzi & al. 2014, Ray 2018, Iphar & al. 2014, Iphar & al. 2015) (Figure 3). En outre, la falsification des messages, c'est-à-dire le fait de volontairement transmettre des informations erronées pour dissimuler par exemple l'activité criminelle d'un navire (pêche dans une zone protégée) (Donati & Fineren 2012), sa route ou tout simplement éteindre le système AIS pour empêcher sa localisation sont d'autres mentions intentionnelles et frauduleuses qui soulignent les limites du système. Toutes ces

manipulations ont pour conséquence de gravement altérer l'intégrité des données et par extension le système en lui-même (Ray & al. 2016, Mazzarella & al. 2017).

Figure 3 : Statistiques sur les fraudes du système AIS



(Windward 2014)

Dans le cas de la non-intentionnalité, il convient de rappeler que les données AIS dépendent de deux « sources » distinctes : d'une part les systèmes embarqués directement connectés à l'AIS (GPS, etc.), et d'autre part de l'intervention humaine non seulement à l'installation du système, mais aussi lors de son utilisation. En ce qui concerne les données issues des différents dispositifs, dans leur étude publiée en 2012 et menée dans la baie de Gdansk en mer Baltique à trois dates différentes et à trois ans d'intervalle, Felski et Jaskolski démontrent que certains champs de données dynamiques comme le ROT[7] ou le HDG[8], contenus dans les messages de type 1,2 et 3, émis par les navires, ont des valeurs incomplètes ou non valides qui oscillent entre 16 % et 23 % (Felski & Jaskolski 2012, p. 19). En ce qui concerne l'intervention humaine sur

[7] ROT : *Rate Of Turn.* Indique le taux de rotation d'un navire en degrés par minute
[8] HDG : *Heading*. Le cap d'un navire est la direction de la boussole dans laquelle la proue est pointée

le système, l'étude d'Abbas Harati-Mokhtari, sur un échantillon « mondial » de donnée AIS, met en évidence que le facteur humain est la cause principale du manque de fiabilité des données AIS. Harati-Mokhtari relève tantôt que plusieurs champs des données, et plus particulièrement les données « statiques », configurées au moment de la mise en marche du système par des techniciens (c.-à-d. *MMSI, type de navire, nom du navire, call sign, caractéristiques du navire*) tantôt que certaines données en lien avec le voyage du navire (*ETA, statut de navigation, destination*) contiennent des erreurs dans la moitié des cas ou alors ne sont tout simplement pas renseignés pour approximativement 8 % des données [9](Harati Mokhtari & al. 2007) (Annexes 6-10).

Illustrons ces propos par le biais d'un exemple concret (Figure 4). En reprenant la nomenclature du message AIS n 5, nous pouvons constater l'hétérogénéité des valeurs contenues dans les messages AIS : des valeurs numériques (n° d'identifiant, données physiques), des valeurs textuelles (nom du navire, destination, etc.), des valeurs prédéfinies, etc. (Iphar & al. 2016, p. 369-370). Par conséquent, non seulement les données introduites dans les systèmes AIS doivent répondre à des normes syntaxiques et de format (car le nombre de caractères par champs de données est limité), mais surtout l'utilisation du système par l'équipage requiert de respecter certaines recommandations internationales (Annexes 4-5). Par exemple, il a été mis en évidence que la destination d'un navire, qui se paramètre au départ par un membre de l'équipage, ne répond pas dans plus de 50 % des cas à la recommandation en matière de nommage des ports selon l'IMO et le UN/LOCODE (Steidel & al. 2019). Dans l'idéal, toute intervention humaine sur le système AIS devrait se faire en connaissance de ces contraintes et selon les usages en vigueur. Pourtant, le système AIS se caractérise par le fait que les communications ne sont ni vérifiées et authentifiées par une quelconque autorité, ni encryptées et donc protégées, menant inévitablement à questionner la qualité et la fiabilité des données. Dès lors, l'appréciation de la véracité ou non des messages AIS, et donc le potentiel de l'information qu'ils contiennent, est étroitement lié à l'évaluation de la qualité des données reçues (Iphar & al. 2015, p. 3).

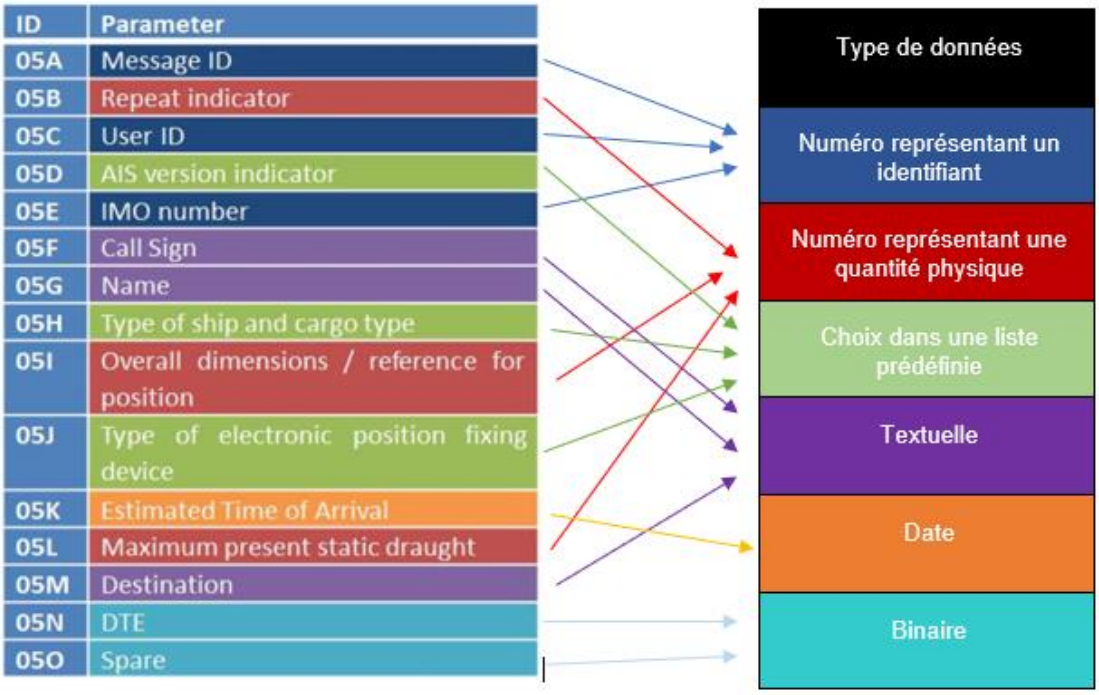
3.3 Cap sur les données du Web public

En l'espace d'une décennie, le système AIS est devenu le principal pourvoyeur de données maritimes dont l'utilité dépasse dorénavant le cadre unique de la sécurité et de la sureté en mer. Peu importe les raisons, qu'elles soient intentionnelles ou non, qui

[9] Dans son analyse, 30'946 messages sur 400'059, soit un ratio de 1/14, contenait au moins 1 erreur (Harati Mokhtari & al. 2007, p. 375)

poussent à remettre en question la fiabilité de ces données, il convient d'être prudent et attentif aux informations que fournit le système AIS.

Figure 4 : Les données et valeurs du message n 5



(Iphar et al. 2016, p. 370)

Un moyen de dépasser cette problématique réside dans l'enrichissement et le croisement des données AIS avec des données d'autres bases de données ou sites internet (Claramunt & al. 2017, Fabbri & al. 2015). Grâce au développement d'outils de Web scraping, il est désormais possible de récolter des données éparses issues de différentes sources dans le but de les compiler et de les exploiter afin d'obtenir une image plus fidèle, plus précise et plus profonde du TMMP (Kazemi & al. 2013, p. 5719).

3.3.1 Le Web scraping

Le Web Scraping se définit comme l'action ou le processus, qui permet de collecter et d'extraire manuellement ou automatiquement, via des scripts ou des programmes, des données en provenance du Web, et notamment de pages HTML (Malik & Rizvi 2011). Généralement, cela implique également de formater les données réunies dans un fichier, de format CSV par exemple. Souvent affilié au Data Mining, le Web Scraping se différencie car il n'implique pas d'analyser les données. De la même manière, le Data Mining n'implique pas l'extraction de données, mais se concentre uniquement sur les processus complexes, pouvant impliquer algorithmes et machine learning, de l'analyse de grands jeux de données. On peut donc considérer que les techniques de Web

Scraping peuvent être utilisées pour créer les jeux de données plus tard utilisées en Data Mining (Parse hub 2020).

Comme le souligne l'étude de Kalyvas & al., les sources d'information ne manquent pas pour effectuer une telle opération : des sites internet payants spécialisés dans le suivi en temps réel des navires (p. ex. MarineTraffic, VesselFinder, etc.) en passant par les données météorologiques, tout est bon à prendre dans l'objectif d'accroître la véracité des données AIS (Kalyvas & al. 2017). Une étude en particulier s'est employée à évaluer plusieurs sources d'informations issues du Web, allant des sites mentionnés précédemment aux sites d'agences internationales et selon leur accès de payant à gratuit, afin de déterminer la qualité de l'information mise à disposition (Strozyna & al. 2018). Cependant, cette étude vise avant tout à démontrer qu'il est nécessaire de procéder à l'évaluation rigoureuse d'une source de données avant de l'exploiter. Elle ne s'attache à aucun moment à l'évaluation effective des données provenant de ces sites. Pourtant les données du Web sont très souvent proposées de manières hétérogènes, peu ou pas structurées ce qui rend difficile non seulement l'appréciation de leur qualité, mais aussi l'intelligibilité des informations qu'elles recèlent et donc des opportunités qu'elles représentent. Ce travail souhaite donc combler ce manque par l'analyse de données de sites de ports de commerce librement accessible sur internet.

3.4 La qualité des données

Au même titre que les données du système AIS sont sujettes à caution, les données récoltées issues du Web public se doivent d'être évaluées. En effet, une telle quantité de données ne peut générer de la valeur que si les décisions et actions qui en découlent reposent sur des données complètes, fiables et correctes. Procéder à l'évaluation de la qualité de ces données permet donc de repérer celles qui sont insignifiantes et non pertinentes (Ardagana & al. 2018, p. 548-549). Cette tâche doit même être considérée comme primordiale afin de s'assurer que seules les données les plus appropriées et les plus dignes de confiance soient disponibles pour l'utilisateur (Hartig & Zhao 2009, p. 9). La notion de qualité des données a été très largement traitée dans la littérature scientifique et se définit le plus souvent comme « *fitness for use* » ou le degré auquel les données satisfont aux besoins de l'utilisateur dans la réalisation d'une tâche spécifique (Wang & Strong 1996, p. 6, Scannapieco & al. 2005, Strozyna & al. 2016, p. 220). Dans cette idée, des données qui seraient incomplètes pour un utilisateur pourraient être tout à fait utilisables pour un autre. En outre, la mauvaise qualité des données pourrait donc être facilement identifiée par exemple par l'absence d'une partie de la donnée ou d'une donnée n'ayant pas la bonne valeur en fonction du contexte d'utilisation.

3.4.1 Data Quality Methodology

La problématique de la qualité des données n'est pas récente : d'abord un champ d'études pour les statisticiens avant d'être une préoccupation pour les chercheurs en management, c'est dans les années 90 que les informaticiens (Fox & al. 1994) se sont intéressés aux moyens de mesurer et d'améliorer la qualité des données contenues dans des bases de données et dans les entrepôts de données (Scannapieco & al. 2005, p. 6). Le développement durant ces dernières années du Web des données n'y a pour ainsi dire rien changé. Au contraire, l'exploitation de plus en plus répandue des données issues du Web public, et par conséquent les biais qu'elles peuvent contenir, n'a en fait que renforcé le besoin d'évaluer la qualité des données (Cai & Zhu, 2015, p. 1). Dès lors, ce changement de paradigme nécessite de mener une réflexion sur les méthodologies d'évaluation de la qualité des données (DQM), et plus particulièrement sur la phase de *Data Quality Assessment* (DQA) jusqu'alors principalement focalisée sur les caractéristiques de la qualité des données issues des bases de données ou des dépôts de données.

Le processus d'évaluation de la qualité des données peut être réalisé de deux façons distinctes : de manière subjective, cette dernière est qualitative par nature et généralement réalisée par le biais de questionnaires ou d'une enquête auprès d'utilisateurs ou d'experts. Ou de manière objective, l'évaluation est conduite par le biais de l'identification préalable de dimensions : celles-ci sont les caractéristiques, attributs ou facettes d'une donnée, permettant ainsi de mesurer la qualité par le biais de métriques pour chacune d'entre elles (Pipino & al. 2002, p. 211, Batini & al. 2009, p. 3). De cette manière, il est possible de comparer et d'évaluer les dimensions, et *in fine* la qualité des données, provenant par exemple de plusieurs sources de données. En comparant les résultats, par rapport à un standard ou un seuil de tolérance, il est possible de juger de la qualité de cette source de données et donc justifier ou rejeter son intégration dans les données propriétaires existantes. Comme nous le verrons plus tard dans ce travail, nous avons conduit un processus hybride mêlant les deux approches. Ceci d'une part pour avoir l'avis d'un expert, cerné les besoins d'un utilisateur mais aussi afin d'avoir des mesures objectives de la qualité des données dans le but de comparaison.

3.4.1.1 DQA pour base de données ou dépôt de données

Dans une étude comparative comprenant une dizaine de méthodologies d'évaluation de la qualité des données préexistantes à l'essor du Big Data, Batini & al. définissent la notion de *Data quality methodology* (DQM) comme un ensemble de lignes directrices et de techniques, qui à partir des informations contextuelles et de l'objectif de l'utilisation

des données, délimitent un processus rationnel d'évaluation (DQA) ayant en point de mire l'amélioration continue de la qualité des données (Batini & al. 2009, p. 1). De cet article, les auteurs mettent en évidence que les méthodologies étudiées répondent toutes à un modèle en trois phases : 1) État de l'existant (*State reconstruction*) consistant à glaner des informations contextuelles sur les processus organisationnels, les collectes de données et les procédures de gestion associées, les problèmes de qualité et les coûts correspondants ; (cette phase peut être ignorée si des informations contextuelles sont disponibles à partir d'analyses précédentes). 2) Évaluation et mesures (*Assessment/measurment*) visant à mesurer la qualité des données grâce aux dimensions de la qualité identifiée comme pertinentes. 3) Amélioration (*improvement*) relative à la sélection d'étapes, de stratégies ou des techniques pour optimiser la qualité des données (Batini & al. 2009, p. 2). Par le biais de cette étude, nous nous concentrons uniquement sur l'étape *Assessment/measurment* car la récolte de données sur le Web représente un moyen d'améliorer les données déjà à disposition suite au constat que ces dernières étaient de mauvaise qualité.

3.4.1.2 Un DQA pour les sources accessibles en open data

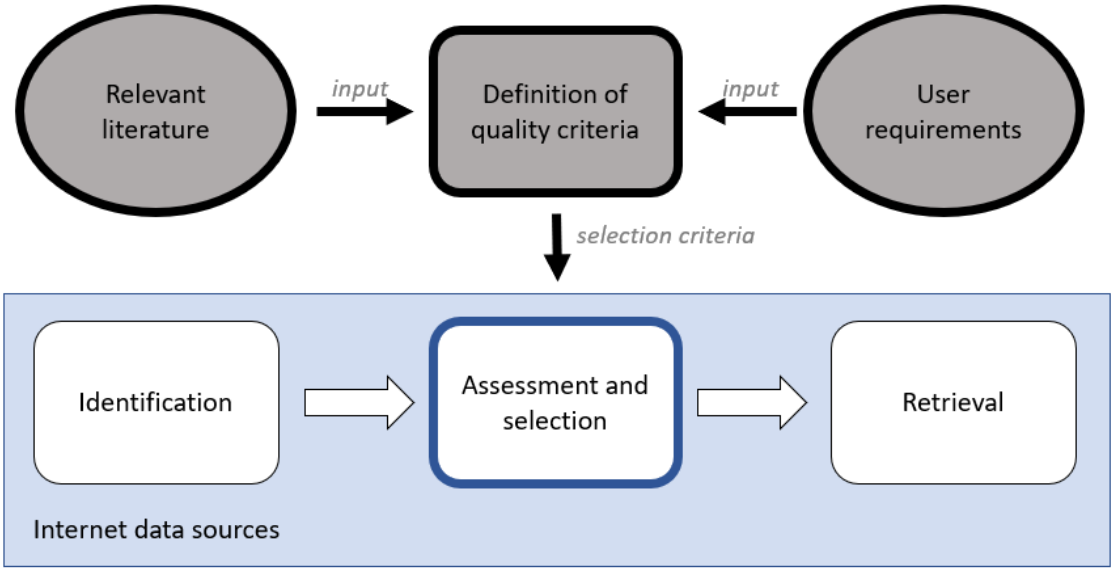
Comme nous l'évoquions précédemment, le développement du Web des données permet aux entreprises de s'appuyer sur une source d'information supplémentaire et/ou complémentaire leur permettant d'optimiser leurs décisions. Partant de ce constat, Strozyna & al. ont imaginé un framework (Figure 5) dans le but d'identifier, d'évaluer et de sélectionner des sources de données issues du Web, et ceci directement en lien avec le domaine maritime, en fonction de certains critères de qualité (Strozyna & al. 2018). Grâce à leur recherche, nous pouvons appréhender les différentes étapes du processus de DQA pour l'évaluation des sources Web avant de décider de leur exploitation ou non. Dans un premier temps, des sources potentielles sont choisies avant d'être évaluées. L'évaluation repose spécifiquement sur la compréhension des besoins de l'utilisateur. Ces derniers déterminent le choix de critères de qualité qui vont ensuite être appliqués pendant la phase d'évaluation. Dans le cas présent, celle-ci s'opère par le biais d'un groupe d'experts attribuant une appréciation de la source en fonction des critères identifiés, ce qui déterminera si cette dernière doit être écartée ou non en la comparant à un seuil de tolérance (*threshold*).

Cette étude met en évidence les enjeux, les problématiques et les étapes à prendre en compte lorsqu'un utilisateur souhaite exploiter des sources Web. En effet, peu importe le type de données, chaque source qui sera utilisée dans un contexte stratégique ou commercial pour une entreprise, doit avoir été évaluée de manière appropriée et dans

le but de s'assurer de la plus haute qualité de l'information pour la prise de décisions (Strozyna & al. 2018, p. 1).

Cependant, l'évaluation des sources par un panel d'experts selon des critères de qualités très précis nous paraît problématique, et ceci surtout dans notre contexte d'étude. En effet, notre travail vise à offrir une méthode implémentable, objective et reproductible qui permet à notre mandant d'évaluer, pour ainsi dire, autant de sources qu'il le souhaite. L'appel à des experts à chaque nouvelle source nous semble donc contradictoire avec l'efficience et l'agilité dont les entreprises doivent faire preuve dans le contexte très concurrentiel du Big Data. De plus, rappelons ici que ce travail vise uniquement à évaluer la qualité des données issues des sites internet de ports commerciaux afin d'enrichir les données déjà à disposition de l'entreprise Riverlake. Néanmoins, et bien que le choix de la source étant déjà arrêté pour notre étude, l'application de ce framework à d'autres sources peut s'avérer instructif pour l'approfondissement de notre champ d'études ou pour de futures recherches.

Figure 5 : A framework for the quality-based selection and retrieval of open data



(Strozyna & al. 2018, p. 225)

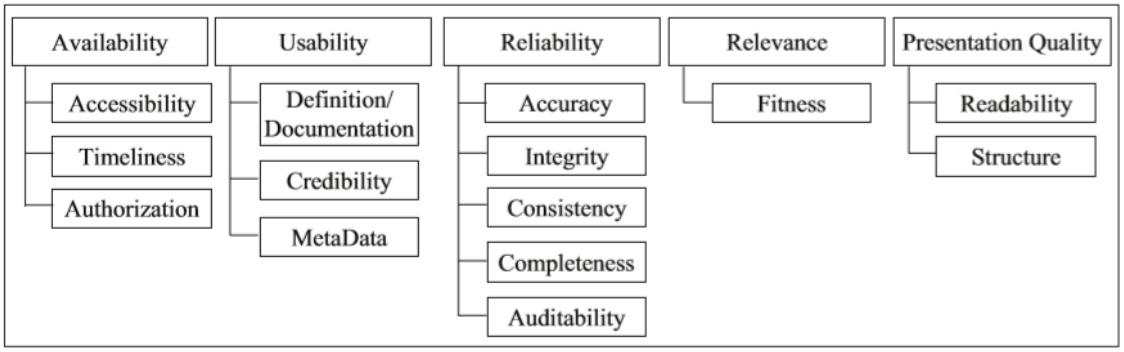
3.4.1.3 Un DQA inspiré du Big Data

En s'appuyant sur l'idée que le Big Data présente des caractéristiques tout à fait spécifiques, s'appréhendant par le biais des 5V[10], Cai & Zhu argumentent que la qualité de ses données fait du même coup face à de nombreux défis (Cai & Zhu 2015, p.2). Pour eux, au vu de la nouveauté du concept de Big Data, il n'existe pas encore de définition uniforme qui saisisse clairement d'une part la qualité de ses données et d'autre

[10] Volume, Vitesse, Variété, Véracité, Valeur

part les critères qui permettent de l'évaluer. Toutefois, comme la majorité des auteurs d'étude sur la qualité des données, ces derniers relèvent l'importance contextuelle de la qualité ainsi que l'importance des besoins des utilisateurs des données. Dès lors. Ils ont imaginé un standard de qualité « universel » reposant sur 5 dimensions principales : *availability,* ou le degré de commodité pour l'utilisateur d'obtenir des données ou des informations sur les données ; *usability,* signifiant que les données sont utiles et répondent aux besoins des utilisateurs ; *reliability,* se référant à déterminer si les données sont dignes de confiance ; *relevance,* est utilisé pour décrire le degré de corrélation entre le contenu des données et les attentes des utilisateurs ; *presentation quality* fait référence à une méthode de description valide pour les données permettant aux utilisateurs de bien les comprendre (Cai & Zhu 2015, p. 4-5). Ces dimensions reposent elles-mêmes sur plusieurs éléments de qualité, dont les auteurs donnent des indicateurs afin de les identifier (p. ex. *accuracy, fitness, credibility*) (Figure 6).

Figure 6 : A universal, two-layer big data quality standard for assessment



(Cai & Zhu 2015, p.4)

Cet article met en évidence l'impact de l'avènement du Big Data et ses conséquences sur l'appréciation de la qualité des données ainsi que sur les nouvelles dimensions et nouveaux éléments à prendre en compte lors de son évaluation. En outre, à l'image de l'article exposé au chapitre précédent, le DQA présenté dans cette étude démontre que l'évaluation de la qualité des données est une étape qui ne peut s'envisager indépendamment des étapes qui la précèdent (p. ex. choix des dimensions de la qualité, choix d'indicateurs, etc.).

En revanche, les auteurs de cet article n'exposent pas de méthode pour procéder à la sélection des dimensions et des éléments qui les composent. Les auteurs soulignent que la phase d'évaluation se découpe à l'heure actuelle en deux catégories : soit qualitative en se basant sur l'opinion et les connaissances d'experts, soit quantitative, grâce à l'obtention de données chiffrées sur les données. Ainsi, les auteurs ne proposent pas des métriques pour procéder effectivement à l'évaluation de leurs dimensions.

3.4.1.4 Un DQA spécifique à notre sujet d'étude ?

L'examen de la littérature à laquelle nous avons pu accéder ne nous permet pas de dégager un DQA sur mesure et prêt à l'emploi par rapport aux données récoltées sur le Web. En revanche, l'agrégation des deux frameworks mentionnés dans le chapitre précédent semble répondre aux objectifs de cette recherche. D'une certaine manière, la méthodologie développée par Strozyna & al. renforce le processus des méthodologies traditionnelles identifiées par Batini & al. En effet, cette dernière se positionne comme une étape supplémentaire à l'appréhension de la qualité des données pour judicieusement choisir la source des données. De ce fait, cela permet de limiter les interventions sur les données et d'éviter ainsi un processus plus long et plus coûteux avant d'extraire de la connaissance/information issue de ces mêmes données (Loshin 2006). De plus, l'idée de faire appel à un expert dont les connaissances permettent de choisir avec plus de pertinence et de rigueur les dimensions de la qualité à évaluer semble être une piste à ne pas négliger lors de l'évaluation.

D'un autre côté, la proposition de framework de Cai & Zhu s'attachant à l'identification de dimensions de la qualité des données nouvellement apparues avec le développement du Big Data finissant ainsi de dresser le contour des caractéristiques nécessaires à notre étude. Cependant, les auteurs ne font aucune mention de la manière pour sélectionner les dimensions, se limitent à celles qui ont été identifiées et n'abordent pas du tout l'objectivation des dimensions par des métriques. Il convient donc d'encore examiner la littérature pour répondre à ces limites.

Néanmoins, l'analyse approfondie des trois articles cités dans cette partie permet de souligner qu'il est largement admis que toute démarche de DQA doit se baser sur une évaluation tenant compte des besoins des utilisateurs et du contexte d'utilisations de ces données. De plus, cette phase d'évaluation doit être appréhendée comme l'une des parties essentielles d'un processus d'amélioration continue de la qualité des données (Batini & al. 2009, p. 2-3, Loshin 2006, p. 11, DAMA 2013, p.5). En effet, l'évaluation doit permettre de mettre à jour les lacunes des données, ainsi que leurs spécificités, dans le but idéal d'en améliorer la qualité de manière continue et d'en monitorer l'évolution dans le temps (cf. 3.4.1.1).

3.5 Les dimensions de la qualité des données

Il n'y a pour ainsi dire pas d'évaluation de la qualité des données sans dimensions, mais leur sélection est peut-être la tâche la plus complexe du processus d'évaluation (Pipino & al. 2002, p.213). Une dimension se comprend comme une caractéristique ou un attribut d'une donnée qui peut être mesurée afin d'en objectiver la qualité. Ceci non

seulement pour déterminer les mesures pertinentes à l'évaluation mais surtout dans le but d'implémenter un contrôle crédible de sa qualité (DAMA 2013, p. 3, Batini & al. 2009, p. 2-3) (Figure 7). Cependant, la notion de qualité des données se perçoit différemment selon les utilisateurs et surtout selon le contexte de l'utilisation des données mettant en évidence la complexité de son évaluation (Wand & Wang 1996, p. 86, Wang & Strong 1996, p. 6, Jesilevska 2007, p. 89-90). Ceci a pour conséquence qu'il n'existe pas de consensus tant sur les dimensions utiles à l'évaluation de la qualité des données que sur leurs définitions comme le démontre le Tableau 1 (Scannapieco & al. 2005, p. 12-13). De plus, il ne faut pas omettre que le développement d'internet et l'impact que cela a provoqué sur les données, illustré notamment par les 5V du Big Data, a fait émerger des dimensions soit qui n'existaient pas auparavant soit qui avaient moins d'importance dans les études en lien avec les bases de données traditionnelles. Pourtant, en s'attardant sur la littérature récente, il est possible de dégager un certain nombre de dimensions qui apparaissent comme essentielles à la qualité des données. C'est ainsi qu'en effectuant un recensement des dimensions les plus souvent citées dans un corpus d'une vingtaine d'études, ce travail met en évidence que certaines dimensions constituent l'épine dorsale d'une DQA à savoir : la complétude (*completeness*), l'unicité (*uniqueness*), la précision (*accuracy*), la validité (*validity*), l'actualisation (*timeliness*) et la cohérence (*consistency*) (Annexe 13).

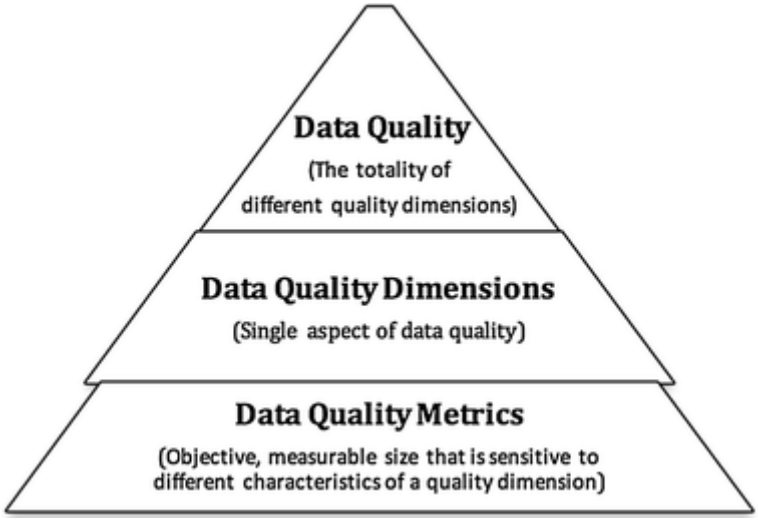
Tableau 1 : Divergence sur la définition de la dimension « cohérence »

Référence	Définition
Pawar (2017)	Implies that not two or more values conflict with each other.
Fox & al. (1994)	Data is said to be consistent with the respect to a set of data model constraints if it satisfies all the constraints in the set.
Iphar & al. (2015)	Coherence of information within a message or between messages
McGilvray (2008)	A measure of the equivalence of information stored or used in various data stores, applications, and systems, and the processes for making data equivalent.
Wang & Strong (1996)	The extent to which data are always presented in the same format and are compatible with previous data

3.5.1 Les métriques issues des dimensions de la qualité des données

Une évaluation objective de la qualité des données repose sur des dimensions qui ont été sélectionnées en fonction du contexte de leur utilisation ainsi que des besoins des utilisateurs. Une fois ces dernières soigneusement choisies, il est indispensable de les associer avec des métriques afin de mettre en lumière la qualité effective des données. Cependant, il convient de souligner que la littérature sur le sujet ne fournit pas un ensemble de métriques clés en main. Au contraire, la plupart des mesures de la qualité des données sont imaginées pour résoudre des problématiques spécifiques à une situation ou une organisation (Cappiello & al. 2004). Comme nous le verrons par la suite, notre étude ne déroge pas à ce constat : pour atteindre ses objectifs, cette dernière doit développer son propre modèle d'évaluation et ses propres mesures.

Figure 7 : Data quality pyramid for successful operationalization



(Gebauer & Windheuser *in* Azeroual & al. 2018)

3.5.1.1 Des mesures non contextuelles

Dans son article *Data Quality Assessment (DQA),* Pipino décrit trois types de métriques (functional forms) utiles à la mesure objective des dimensions : 1) Ratio simple, qui peut par exemple exprimer le nombre de résultats désirables par rapport au nombre total des résultats ou inversement. 2) Opération Min ou Max, qui est utilisé lorsque plusieurs indicateurs sont pris en compte pour la valeur d'une dimension. L'opérateur Min agit de manière conservatrice et sélectionne la valeur la plus basse. D'un autre côté, l'opérateur Max agit de manière optimiste et sélectionne la valeur la plus élevée. 3) Moyenne pondérée, qui est une alternative au fonctionnement Min ou Max. Dans cette mesure, chacun des indicateurs se voit attribuer un « poids » pour déterminer l'importance de ce dernier dans la valeur finale de la dimension, puis une moyenne pondérée est calculée. Les valeurs de poids sont comprises entre zéro et un, et leur valeur totale est égal à un, afin d'obtenir une note normalisée. En outre, il est mentionné que ces mesures sont soit

indépendantes de la tâche (*task-independent*), soit dépendantes de la tâche (*task dependent*). Les premières ne prennent pas en compte le contexte dans lequel la mesure est effectuée au contraire des deuxièmes qui s'appuient sur le contexte. (Pipino & al. 2002). Pourtant, cette étude ne présente pas d'étude de cas approfondie sur ces mesures de manière générale et sur la distinction qui s'opère entre les mesures *task-independent* et *task-dependent*. À l'image des autres méthodologies d'évaluation de la qualité des données, cet article se borne à exposer des mesures standards, parfois plus ou moins complexe, pouvant s'appliquer à une multitude de situations et de cas sans approfondir l'impact du contexte sur ces dernières.

3.5.1.2 La mesure de la qualité par le biais de métriques pondérées

Dans le cadre de cette étude, il est nécessaire d'identifier des mesures qui tiennent compte des spécificités contextuelles dans lesquelles elle se déploie. Pour ce faire, l'utilisation de métriques pondérées s'avère très utile (Vaziri & al. 2019, Wei & al. 2016 p. 10). Cette idée se base sur le concept selon lequel, dans une entreprise, certaines données peuvent être plus importantes et pertinentes à un moment particulier ou tout au moins dans un contexte commercial précis. Par conséquent, ces données doivent recevoir plus de poids dans les mesures en fonction de la finalité poursuivie. Dans le cas de Riverlake et du TMMP, les données dites de navigation (ETA, destination etc.) s'avèrent plus utiles que les données statiques (n° IMO, pavillon du navire, etc.) et donc être évaluées en conséquence. Ceci peut également être vrai pour les données non significatives dont le poids sera moindre ou nul dans les mesures (Vaziri & al. 2019, p. 709).

3.5.1.2.1 Exemple par rapport à la complétude des données

Reprenons ici un exemple, transposé au trafic maritime, largement inspiré du travail de Vaziri & al. afin de mieux comprendre l'apport de la pondération des données (Vaziri & al. 2019, p. 713-714). La complétude peut être définie comme la mesure des données non manquantes. Dans le tableau 2, des valeurs fictives quant à la complétude d'une base de données AIS sont représentées. Le taux de complétude correspond au ratio du nombre de valeurs connues par rapport au nombre de valeurs total dans une colonne. Par exemple, pour la colonne *Nom*, nous connaissons 90 valeurs sur 100 possibles.

Tableau 2 : Échantillon fictif du taux de complétude par colonnes

Nom	Pavillon	N° IMO	Vitesse	Destination	ETA
90 %	95 %	75 %	50 %	25 %	30 %

Le taux moyen de complétude de ce jeu de données, ici **60,83 %**, se calcule par l'addition du taux de toutes les colonnes (365) divisé par le nombre de colonnes (6) (tableau 2).

(90 +95 +75 +50 +25 +30)/6 = 60,83

Cependant, si un poids est défini pour chaque colonne du jeu de données en fonction de la tâche à accomplir ou des besoins de l'utilisateur, il alors est possible de calculer une valeur plus représentative de la qualité des données par rapport au contexte d'utilisation (*task-dependent*). Par exemple, il est possible d'imaginer que le nom d'un navire, sa destination ainsi que son ETA soient les informations les plus importantes dans un jeu de données et dans un contexte spécifique. Le tableau 3 contient des poids s'inspirant de notre contexte d'utilisation.

Tableau 3 : Échantillon fictif du poids par colonnes

Nom	Pavillon	N° IMO	Vitesse	Destination	ETA
0,25	0,10	0,10	0,15	0,20	0,20

Dans ce cas précis, le taux moyen de complétude pondéré serait alors de **58 %**, soit l'addition du taux de complétude de chaque colonne multipliée par leur poids respectif (tableau 4).

(90*0,25) + (95*0,10) + (75*0,10) + (50*0,15) + (25*0,20) + (30*0,20) = 58

Ce modèle est intéressant car il est ensuite possible d'appliquer ces calculs à la plupart des dimensions permettant alors d'obtenir une synthèse chiffrée globale de la qualité en fonction d'un contexte d'utilisation déterminé (tableau 4). Le résultat ainsi obtenu permet donc non seulement de connaître la qualité générale d'une base de données mais également la comparer avec d'autres bases.

Tableau 4 : Mesure des dimensions et de leurs poids

Dimension	Mesure	Poids
Complétude	**0,58**	0,4
Précision	0,62	0,1
Cohérence	0,75	0,3
Validité	0,50	0,2

(0,58*0,4) + (0,62*0,1) + (0,75*0,3) + (0,5*0,2) = 62

Dans cet exemple, le coefficient de qualité de cette base de données est alors de 62 %, ayant pris en compte les poids pondérés des colonnes les plus importantes pour l'utilisateur, mais aussi la pondération des dimensions selon ses besoins.

3.5.1.3 L'apport des mesures à l'amélioration de la qualité

Bien que ce travail se concentre à l'évaluation de la qualité des données provenant des ports commerciaux, nous souhaitons relever ici que les mesures effectuées sur les dimensions de la qualité permettent également de monitorer l'évolution de la qualité des données d'une source ou d'un jeu de données dans le temps. Celles-ci objectivent non seulement la notion de qualité, mais servent également d'indicateur dans le but d'identifier les erreurs/manques, de les gérer et de contribuer à l'amélioration des données à disposition pour l'accomplissement d'une tâche. De sorte que les *business policies* fixées par exemple par une entreprise quant à la qualité des données utiles à son activité, par le biais de la sélection de dimensions et de métriques pertinentes, lui permette de s'assurer de la valeur de ses données pour répondre au mieux à ses objectifs commerciaux tout en mitigeant ses risques (Loshin 2006, p. 3).

3.6 Conclusion

Cet état de l'art de la littérature consacrée d'une part au système AIS et d'autre part à la qualité des données permet de mettre en évidence les principaux enjeux de la problématique.

Après avoir présenté les raisons d'être et le fonctionnement du système AIS, nous avons brièvement abordé les études qui se consacrent désormais à l'exploitation des données de ce dernier démontrant ainsi tout son intérêt, et ceci dans plusieurs champs disciplinaires. Cependant, une abondante littérature scientifique nous met en garde quant à la fiabilité et l'intégrité des données issues du système. La principale cause des erreurs constatées dans les données est à attribuer au facteur humain. Car non seulement l'intervention humaine dans l'installation, le paramétrage ou le renseignement des champs libres dans le système par exemple au départ d'un navire, que les tentatives frauduleuses et malintentionnées, sont autant de raisons altérant la confiance dans le système.

Pourtant, l'utilisation d'autres sources d'informations librement accessible sur internet afin de venir compléter, vérifier ou enrichir les données AIS représente une réelle opportunité d'amélioration. Mais à l'image des données AIS, ces dernières doivent être évaluées afin de juger de leur qualité et de leur pertinence avant de contribuer à toute prise de décision commerciale ou stratégique. Cette évaluation repose sur des

dimensions de la qualité sur lesquelles il n'existe, à l'heure actuelle, aucun consensus et dont les définitions sont très variables. Les chercheurs s'accordent pourtant à relever que le contexte d'utilisation des données et les besoins des utilisateurs sont la pierre angulaire du processus d'évaluation. Ce dernier peut être réalisé de manière subjective par la mobilisation d'experts et au travers d'entretien ou de questionnaire. Ou alors il peut être conduit de manière objective par le biais de l'identification de métriques découlant directement des dimensions de la qualité. En outre, l'influence du contexte et des besoins des utilisateurs des données ne sont pas à négliger lorsqu'il s'agit de réaliser l'évaluation finale et ceci peut être pris en compte par l'intermédiaire de métriques pondérées.

4. Méthodologie

4.1 Approche méthodologique générale

Pour rappel, notre étude vise à répondre aux questions de recherche suivante :

- Quelles sont les métriques pertinentes quant à l'évaluation de la qualité des données récoltées dans le Web public ?
- Quel est l'impact des données récoltées par le biais d'outils de Web scraping sur la qualité des données déjà à disposition ?

Ce travail de recherche présente trois caractéristiques essentielles. Dans un premier temps, il s'ancre dans une démarche visant à comprendre l'état actuel de la recherche. Sa visée est donc clairement de type descriptif car elle a « pour but d'obtenir des informations précises sur les caractéristiques à l'intérieur d'un domaine particulier et de dresser un portrait de la situation » (Fortin & Gagnon, 2016, p. 208). De plus, ayant pour objectif de décrire le phénomène de l'exploitation de données issues du Web public dans le contexte du TMMP et d'explorer l'impact de ces dernières sur la qualité de l'information des données déjà à disposition, cette étude est hautement exploratoire. En effet, le type d'étude exploratoire s'intéresse principalement à des phénomènes nouveaux qui ne sont pas ou très peu documentés. En l'occurrence, l'analyse qualitative de données provenant de sites internet de ports commerciaux en Europe et la proposition d'une méthodologie d'évaluation n'a pour l'instant pas bénéficié d'une attention particulière. Cette dimension exploratoire vise donc à se familiariser avec des faits, des situations et des données en lien avec le trafic maritime afin de formuler des questions pour de futures recherches et à générer de nouvelles idées ou hypothèses (Dufour & Larivière 2017).

Deuxièmement, ce travail revêt des traits tant qualitatifs que quantitatifs. Se basant sur un état de la littérature couvrant une grande diversité d'articles scientifiques pour identifier les métriques utiles à l'évaluation de la qualité des données ainsi que sur un

questionnaire visant à mettre à jour les attentes et les besoins du mandant de ce projet, la première phase de ce travail est résolument qualitative (Fortin & Gagnon, 2016, p. 16). Cependant, l'étape suivante, visant à développer une méthodologie d'évaluation grâce à l'objectivation de la qualité des données basée sur des variables calculées à partir de métriques, propose un indice de qualité reposant sur l'échantillon provenant des données récoltées grâce au Web scraping, s'avère quant à elle plus quantitative.

Enfin, comme l'un des buts poursuivis de cette recherche est de prendre pied sur l'existant afin de proposer certaines recommandations par rapport à l'exploitation de sources internet dans le cadre du TMMP, cela induit logiquement la possible extension de ce procédé à d'autres sources, offrant une dimension supplémentaire à ce travail.

4.1.1 Méthode de recherche

La méthode de recherche de notre étude est l'enquête. Ce devis permettant d'examiner non seulement les besoins et les attentes du mandant quant à la qualité des données par le biais d'un questionnaire, mais également « de recueillir de l'information factuelle sur un phénomène existant, de décrire des problèmes, d'apprécier des pratiques courantes et de faire des comparaisons et des évaluations » (Fortin & Gagnon 2016, p. 211). De plus, il convient de mentionner encore que cette recherche est de type transversal puisqu'elle se base uniquement sur un échantillon de données récoltées durant la dernière semaine de juin 2020. (Fortin & Gagnon 2016, p. 215)

4.2 Collecte de données

Pour cette étude, deux méthodes de collecte distinctes ont été appliquées : dans un premier temps, cette dernière s'appuie sur les données issues de 11 ports commerciaux situés en Europe, récoltées dans le cadre du travail partenaire intitulé « Étude, conception, collecte, curation et évaluation d'un scraping de sites Web liés au transport maritime pour améliorer la prédiction du fret de matières premières » (tableau 5) (Druey 2020). Dans un deuxième temps, non seulement dans le but de cerner les lacunes actuelles, mais aussi dans le but de clarifier les besoins et les attentes de l'entreprise Riverlake en matière de données, un questionnaire a été utilisé (Annexe 14). Ainsi notre mandant a été interrogé au titre d'utilisateur des données autant qu'expert en matière de TMMP.

Tableau 5 : Ports sélectionnés dans la cadre de cette recherche

Port	Pays
Aarhus	Danemark
Amsterdam	Pays-Bas
Bordeaux	France
Copenhague	Danemark
Dunkerque	France
Fredericia	Danemark
Hambourg	Allemagne
Klaipeda	Lituanie
Le Havre	France
Niedersachsen	Allemagne
Rotterdam	Pays-Bas

(Druey 2020)

4.2.1 Questionnaire

4.2.1.1 Choix de la méthode de collecte

Conçu en s'appuyant sur les lectures ayant pour sujet la qualité des données et les dimensions nécessaires au processus d'évaluation, ce questionnaire a été mis à disposition par le biais du site de sondage en ligne *LimeSurvey*. Il se compose de 24 questions selon les modalités suivantes (Annexe 14) :

- 22 questions préformées en lien avec les dimensions utiles à l'évaluation de la qualité des données pour lesquelles un choix est imposé parmi 5 possibilités
- 1 question proposant au participant de classer selon son appréciation ces mêmes dimensions par ordre d'importance
- 1 champ libre offrant au participant la possibilité de laisser une ou plusieurs remarques afin d'éventuellement apporter des précisions à ses réponses

4.2.1.2 Population cible

Notre état de la littérature nous a permis de mettre en évidence que l'évaluation de la qualité des données repose essentiellement sur deux points cruciaux : le contexte d'utilisation des données ainsi que les besoins de l'utilisateur des données. Dans cette idée, il était donc nécessaire d'interroger le mandant de cette étude quant aux points évoqués. En effet, d'une part il connaît précisément l'état actuel des données à sa disposition et les attentes qu'il a par rapport aux données récoltées sur les sites sélectionnés pour cette recherche, mais d'autre part, son expertise du TMMP nous est très utile pour comprendre le contexte d'utilisation des données et donc d'appréhender

la qualité nécessaire à la réalisation de ses objectifs. Par conséquent, ce questionnaire et sa passation se limitent à notre seul mandant.

4.2.1.3 Rédaction du questionnaire

Le questionnaire transmis à notre mandant vise trois objectifs principaux :

1. Sélectionner les dimensions qui s'avèrent pertinentes dans le contexte du TMMP par rapport à celles qui sont citées dans la littérature scientifique
2. Obtenir un état actuel des données à disposition de Riverlake
3. Cerner et objectiver les attentes quant aux données obtenues par le biais du Web scraping

4.2.1.3.1 Sélection des dimensions

Comme cela a déjà été mentionné dans ce travail, les dimensions de la qualité des données servent à objectiver la notion de qualité dans le but d'en dégager les métriques utiles à son évaluation. Pour ce faire, 20 articles scientifiques traitant de la qualité des données et des dimensions de la qualité des données ont été sélectionnés (Annexe 12). Cette sélection repose sur deux critères principaux :

1. Elle comporte des articles traitant tant des dimensions de la qualité des données en lien avec des bases de données traditionnelles que des études plus récentes quant à ces mêmes dimensions dans le contexte du Big Data. En effet, l'avènement de ce dernier a poussé à reconsidérer les dimensions traditionnelles de la qualité des données notamment à la vue, par exemple, de l'accessibilité et de l'interprétabilité des données, de la crédibilité de la source utilisée ou encore de l'expressivité de l'information.
2. Afin de s'assurer de la représentativité des études scientifiques sur le sujet et des dimensions de la qualité abordées dans ceux-ci, seuls les articles définissant ou explicitant clairement les dimensions utilisées ont été choisis. Cela implique donc que les études ne définissant pas les dimensions de la qualité ont été écartées. De plus, la sélection a été stoppée lorsque les définitions des dimensions des données sont arrivées à saturation c'est-à-dire lorsqu'aucune nouvelle dimension ou définition n'était proposée.

Dans un deuxième temps, l'objectif a été de compiler les dimensions présentées dans la littérature scientifique ainsi que leurs définitions (Annexe 11). De ce travail, résulte un regroupement des définitions sous le terme de « dimension agrégée ». Par exemple, DAMA définit la complétude comme « the proportion of stored data against the potential of 100% complete » (DAMA 2013). De son côté, Gitzel propose simplement la notion de « not missing data » (Gitzel & al. 2015). Enfin Fox décrit la complétude comme « the degree to which a data collection has values of all attributes of all entities that are supposed to have values » (Fox & al. 1994). Par conséquent, ces trois définitions de la complétude ont été regroupées puisqu'elles décrivent une même opinion sur les données.

Grâce à cette étape, il est ensuite possible de connaître la somme des fréquences de citation de chacune des dimensions. La fréquence signifie donc le nombre de fois qu'une dimension a été citée ou définie dans un article scientifique. Cette valeur est alors employée pour apprécier le niveau d'importance que la littérature scientifique donne à chacune des dimensions. Bien évidemment, cette dernière n'a été comptabilisée qu'une fois par article et les résultats se trouvent en annexe, classés par ordre décroissant de fréquence (Annexe 13).

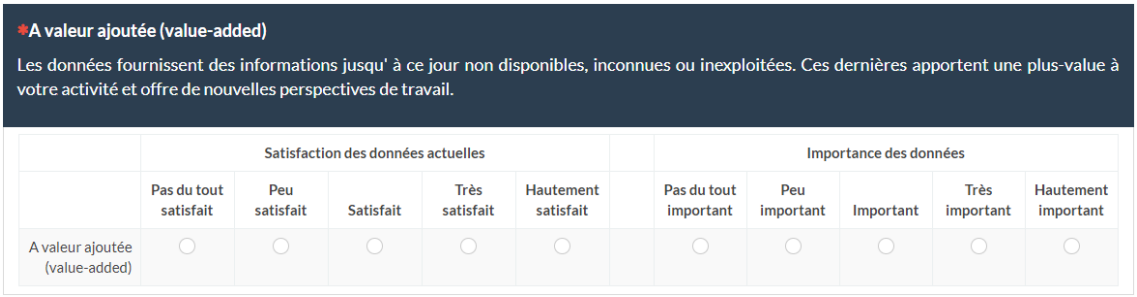
Ce travail a également servi à proposer une définition unique pour chaque dimension. En effet, comme nous l'évoquions déjà, il n'y a pas de consensus sur les définitions des dimensions de la qualité des données. Cependant, sans être d'accord sur la forme, beaucoup d'auteurs sont d'accord sur le fond. À cet égard, le regroupement des dimensions en dimensions agrégées a fourni l'opportunité de définir les dimensions, prenant en compte les diverses caractéristiques avancées par les chercheurs et ceci, dans le but de les proposer dans le questionnaire.

4.2.1.3.2 Appréhender l'état actuel et cerner les attentes

Le travail sur les dimensions effectué, celui-ci a été utilisé dans le but de comprendre l'état actuel des données, mais également de cerner les attentes et les besoins du mandat de cette recherche. En effet, toutes les étapes visant à améliorer des données et à proposer une solution adéquate nécessitent de s'interroger sur l'existant. C'est donc par le biais de ce questionnaire que cette étude s'est affairée à comprendre les lacunes des données actuellement à disposition. Mais c'est aussi par celui-ci que des informations ont été collectées dans l'objectif de cerner concrètement les attentes de l'utilisateur final des données récoltées sur internet.

Pour cela, le questionnaire contient les 22 définitions des dimensions agrégées provenant de la littérature. Dans un souci de clarté, et surtout pour éviter toute redondance ainsi qu'une passation du questionnaire plus aisée, les dimensions et les définitions ne sont explicitées qu'une seule fois. Ainsi, notre mandant évalue à la fois la satisfaction qu'il a des données actuelles et, par la même occasion, il répond en tant qu'expert quant à l'importance qu'il accorde à ces mêmes dimensions dans le contexte du TMMP. Chaque dimension est évaluée selon l'échelle de Lickert offrant ainsi 5 choix au participant (Figure 8). Cette dernière est particulièrement simple à utiliser, ne nécessite aucune connaissance spécifique pour le participant et permet à l'administrateur du questionnaire de forcer un choix unique dans les réponses.

Figure 8 : Exemple de question utilisée pour cette étude



Pour terminer, ce questionnaire permet au participant de classer les 22 dimensions selon leur ordre d'importance. Ceci, dans le but de mettre en évidence les dimensions les plus importantes au regard de son champ d'activité.

Avant de diffuser le questionnaire, nous l'avons fait tester auprès de deux personnes n'ayant aucun lien avec le contexte de cette étude. Ce prétest nous a permis de clarifier d'une part les objectifs du questionnaire et d'autre part nous assurer que les définitions étaient claires et compréhensibles. Finalement, la diffusion du questionnaire a été validée par Arnaud Gaudinat, professeur responsable de la supervision de ce travail de recherche.

4.2.1.4 Limites

4.2.1.4.1 Dans la sélection des dimensions

Une des limites de notre méthodologie concerne l'identification des dimensions et se rapporte directement la littérature scientifique. En effet, la sélection d'articles pour cette étude a utilisé les thèmes/mots clés, seuls ou en les combinant, tels que *Data Quality, Data Quality Assessment, Data Quality Metrics, Data Quality Dimensions, Big Data, Data Quality Measurement.* Bien qu'une grande représentativité a pu être atteinte, il ne s'agit pas pour autant de l'exhaustivité du sujet. Un plus grand nombre de références pourrait probablement apporter encore d'autres dimensions permettant d'améliorer la représentativité. Cependant, nous avons déjà remarqué une saturation au niveau des définitions et des dimensions dans la littérature. Si bien que la sélection des articles s'est arrêtée lorsque les dimensions proposées dans de nouvelles références faisaient appel à des recherches déjà analysées dans le cadre de notre étude.

4.2.1.4.2 Dans les dimensions agrégées et les définitions

Une autre limite à signaler concerne les dimensions agrégées. Ce travail de comparaison et de regroupement a été effectué par l'auteur de cet essai sans méthode et validation particulière. Le but principal étant d'identifier les dimensions les plus citées et utilisées dans la littérature scientifique afin de s'en servir dans le cadre de cette recherche. Le même constat s'impose quant aux définitions de ces mêmes dimensions.

Le questionnaire imaginé dans cette recherche était avant tout destiné au mandant de cette étude afin de comprendre l'état actuel de ses données (satisfaction), ses attentes et besoins (importance) par rapport aux données récoltées mais aussi, d'utiliser ses connaissances du TMMP afin d'identifier les dimensions importantes pour notre champ d'études. S'il est possible d'imaginer que les premiers points le concernent presque exclusivement, le deuxième objectif du questionnaire pourrait bénéficier de plus de représentativité avec la passation d'autres individus. En effet, l'interrogation de plusieurs experts pourrait mettre en évidence d'autres besoins et attentes quant aux données du TMMP.

4.2.2 Web scraping

4.2.2.1 Choix de la méthode de collecte

Il convient de rappeler ici que le présent travail de recherche repose en grande partie sur une autre étude menée simultanément intitulée « Étude, conception, collecte, curation et évaluation d'un scraping de sites Web liés au transport maritime pour améliorer la prédiction du fret de matières premières ». L'objectif de cette recherche vise à appliquer une méthode de Web scraping sur les sites internet de ports commerciaux afin d'y récupérer des données liées au TMMP. Basé sur un choix raisonné, prenant en compte la disponibilité des données, l'importance du trafic commercial de matières premières, l'intérêt du mandant et du projet pour certains ports, onze ports ont été sélectionnés (tableau 5) (Druey 2020).

Dans la mesure où ce scraping permet de récolter une très grande quantité d'informations, souvent très disparates et peu homogènes par exemple, à cause de la langue ou d'abréviations différentes, Guy Druey a mis en place un référentiel qui met en lumière la typologie des données disponibles. Ce dernier a par la suite servi à normaliser les données récupérées pour chaque port. Ainsi, en se calquant sur les types de données présents dans les messages AIS, trois catégories de données ont été créées (Annexe 16) :

- Les données liées à l'identification des navires, dites statiques
- Les données dynamiques directement liées au voyage des navires
- Les données relatives aux caractéristiques techniques des navires

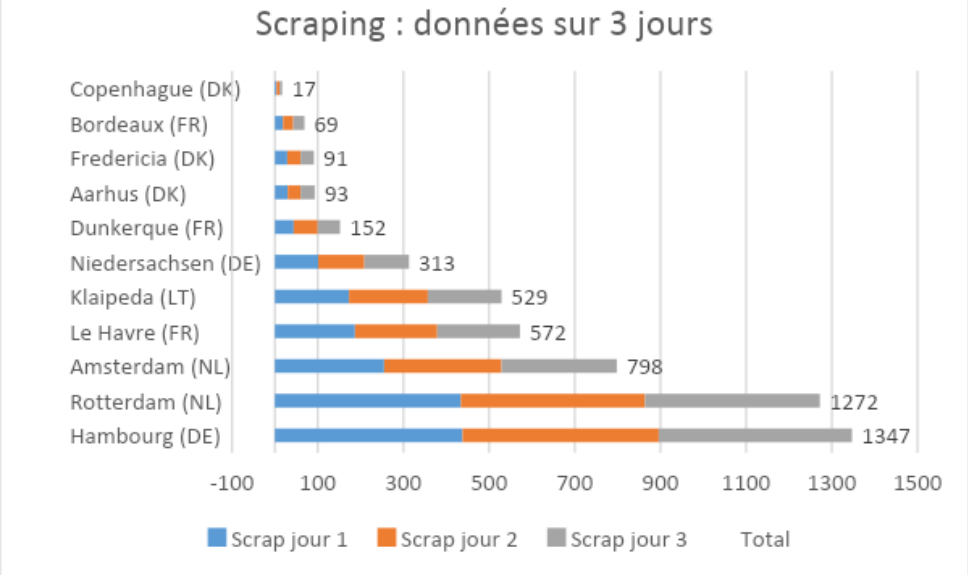
Néanmoins, il convient de réaffirmer que le but du projet repose sur la volonté d'enrichir ou de vérifier des données issues du système AIS, et tout cela en lien avec les éléments de données en rapport avec le voyage des navires. Dès lors, ce sont les données dynamiques qui revêtent le plus d'importance pour le projet global (les deux études partenaires) et notre étude en particulier. Il va de soi que ces données ne sont pas suffisantes car des données statiques doivent y être rattachées pour identifier

correctement les navires concernés par le scraping (IMO, pavillon, nom). Pour terminer, les données liées aux caractéristiques techniques des navires n'étant pas pertinentes dans le cadre de notre étude, ces dernières n'ont pas été retenues pour le référentiel. Ces éléments clarifiés, ce scraping, réalisé entre le 25 et le 27 juin 2020, a permis de récupérer 5253 enregistrements de navires (Figure 9) (Druey 2020).

4.2.2.1.1 Nettoyage des données et uniformisation des champs

À la suite de la récolte de ces données, le chercheur a ensuite procédé au nettoyage de ces dernières et plus particulièrement à l'élimination des doublons, c'est-à-dire la suppression des enregistrements du 25 juin présent dans le scraping du 26 et 27 juin. Ceci a un pour conséquence de faire passer le nombre d'enregistrements de navire de 5253 à 3322. En outre, les champs « statut » et « ETA » « ETD » ont été uniformisés directement dans les scripts du scraping, et donc modifiant en partie les données, dans le but d'avoir des valeurs uniformisées pour tous les ports. Cette opération d'uniformisation a également été conduite pour les champs « pavillon », « type de bateau », « provenance » et « destination » par le biais du logiciel *OpenRefine.*

Figure 9 : Quantité des données scrapées du 25 au 27 juin 2020



(Druey 2020)

4.2.2.2 Échantillon

C'est donc sur la base de données créée ex nihilo par Guy Druey que nous avons ensuite réalisé nos analyses. Pour cela, nous avons réalisé un échantillon aléatoire stratifié proportionnel. Dans un premier temps, nous avons regroupé les ports en strates selon la sélection raisonnée et le regroupement des ports choisis pour le scraping[11] (tableau 6)

[11] Ces strates se basent sur l'état d'avancement du travail de Guy Druey et selon les discussions qu'il a mené avec le mandant et le superviseur au 02.07.2020.

(Druey 2020). Ainsi, cet échantillonnage permet de réduire la taille requise pour obtenir un échantillon représentatif tout en obtenant des strates homogènes et au prorata de la population constituée des 3322 enregistrements (Fortin et Gagnon, 2016, p. 267).

Tableau 6 : Effectifs des strates

	Strate			
Port	**1**[12]	**2**[13]	**3**[14]	**4**[15]
Amsterdam	521			
Hambourg	567			
Rotterdam	984			
Bordeaux		35		
Dunkerque		103		
Klaipeda		266		
Le Havre		405		
Aarhus			56	
Copenhagen			16	
Fredericia			58	
Niedersachsen				311
Total	**2072**	**809**	**130**	**311**

Nous avons ensuite déterminé la taille de l'échantillon nécessaire à la réalisation de notre étude afin d'être représentatif de la population de 3322 enregistrements de navire. Nous avons choisi un niveau de confiance de 95 % pour obtenir un résultat de 345 enregistrements. Dans un troisième temps, nous avons calculé la proportion de chaque strate contenue dans la population totale. Ce résultat a par la suite été multiplié par la taille de l'échantillon minimum nécessaire afin de constituer notre échantillon final (tableau 7).

[12] Strate 1 : ports les plus importants de notre étude en termes de volume de marchandises transbordées

[13] Strate 2 : ports de taille similaire

[14] Strate 3 : ports danois

[15] Strate 4 : port particulier, « Niedersachsen » regroupe 5 ports maritimes et notamment Wilhelmshaven, 3ème port de commerce d'Allemagne transborde 80 % du pétrole brut en Allemagne

Tableau 7 : Nombre d'enregistrements de navires nécessaires par strate

Strate	Effectifs	Proportion	Échantillon	Nbr de navires par strate
1	2072	0,6237		215
2	809	0,2435		84
3	130	0,0391		14
4	311	0,093		32
Total	3322	1	345	**345**

4.2.2.2.1 Choix aléatoire

Afin de sélectionner aléatoirement les enregistrements pour constituer notre échantillon final, nous avons procédé en trois étapes :

1. Regroupement des ports selon leurs strates respectives
2. Attribution d'un numéro aléatoire à tous les enregistrements
3. Classement des enregistrements en fonction d'un numéro aléatoire attribué pour finaliser la sélection

Pour réaliser la deuxième étape, nous avons fait appel à la fonction =ALEA () dans Excel, afin d'attribuer aléatoirement un numéro à chaque entité que nous avons ensuite sélectionnée en fonction de ce même numéro et du nombre d'entités requises pour notre échantillon.

4.2.2.3 Limites

4.2.2.3.1 Le choix des ports

Le choix des ports a été réalisé dans le travail partenaire à cette étude et s'est porté sur des infrastructures situées en Europe. Pourtant comme le relève Guy Druey, plus de la moitié des 20 ports commerciaux européens les plus importants ne fournissent pas de données sur le Web public (Druey 2020). Ainsi les ports d'Anvers, Marseille, Algeciras, Barcelone, Gênes, le Pirée et la totalité des ports situés en Turquie ne proposaient pas de données exploitables au moment du lancement des projets d'études (avril 2020). Depuis, il est possible de récupérer certaines données pour le port de Marseille. En revanche pour les autres cas, différentes raisons techniques ou simplement d'accès à l'information ont limité le choix des ports disponibles (p. ex. la nécessité de créer un

compte). L'accessibilité des données s'est donc avérée cruciale dans la sélection effectuée.

4.2.2.3.2 Les agents maritimes

Dans un premier temps, il avait été envisagé de récupérer des données concernant les agents maritimes. Bien qu'il y ait été possible d'inclure un champ supplémentaire dans la base de données issues des données récoltées pour les agents, celui-ci s'avère très imprécis et sans réel apport d'information. En effet, on peut certes constater la présence d'agents, mais aussi d'affréteurs ou d'opérateurs, ce qui souligne l'imprécision de ce champ. De plus, après investigation, il s'avère que les sites internet des agents maritimes ne fournissent pas de données publiques qui puissent venir compléter ou enrichir les données des ports déjà sélectionnés (Druey 2020).

4.2.2.3.3 Une récolte de données limitée dans le temps

La récolte des données a eu lieu dans un intervalle assez court, du 25 au 27 juin 2020, permettant de récolter des données sur pas moins de 3322 enregistrements. Bien que les données récoltées nous autorisent à déjà émettre certaines hypothèses quant à la qualité des données issues des ports, des données récoltées sur un laps de temps plus important pourraient fournir encore plus de connaissance par rapport à notre champ de recherche. En effet, des données récoltées sur un temps plus long s'avèreraient par exemple très utiles pour la création d'un benchmark de la qualité plus représentatif de la réalité.

5. Présentation et discussion des résultats

5.1 Questions de recherche n 1 : les métriques pertinentes

Notre première question de recherche visait à déterminer les métriques pertinentes quant à l'évaluation des données dans le cadre du TMMP. Pour répondre à cette dernière, nous avons procédé à une revue de littérature approfondie. Nous avons sélectionné 20 références et mis en évidence 22 dimensions. Pour nous assurer de l'adéquation de ses dimensions avec la tâche à réaliser et le contexte de notre étude, nous avons demandé au mandant de cette recherche, interrogé au titre de participant, mais aussi d'expert, d'évaluer ces dernières. La réponse à cette question de recherche s'appuie largement sur les étapes mises en évidence dans les DQA au chapitre 3.4.1.2.

5.1.1 Les dimensions avant les métriques

La première phase de ce travail s'est focalisée sur l'identification d'articles scientifiques permettant de comprendre comment procéder à l'évaluation de la qualité des données. Par le biais des mots-clefs tels que *Data Quality, Data Quality Assessment, Data Quality*

Metrics ou encore *Data Quality Measurement,* l'état de la littérature a mis en évidence que le choix de métriques pertinentes pour évaluer la qualité des données doit se faire par l'identification de dimensions de la qualité des données. En effet, la notion de la qualité et son évaluation objective donc s'appuyant sur des métriques, ne peut s'effectuer sans avoir préalablement sélectionné des dimensions qui rendent compte des caractéristiques des données étudiées. Pour cela, nous avons attentivement examiné 20 références traitant d'une part de la qualité des données « traditionnelles » et des dimensions qui s'y rattachent, c'est-à-dire se concentrant sur les bases de données, et d'autre part sur la qualité des données et les dimensions mises en évidence avec l'avènement du Big Data en toile de fond.

5.1.2 Les dimensions identifiées dans la littérature

Ce travail a permis d'identifier 22 dimensions pour un total de 143 fréquences. Ces fréquences ont ensuite permis d'établir un classement des dimensions selon leur nombre de citations dans la littérature (Annexe 13).

Ce classement démontre que 4 dimensions se détachent clairement des autres avec plus de 10 citations/références dans la littérature (tableau 8). Bien que ce tableau nous permette de mettre au jour l'importance accordée par la littérature aux dimensions de la qualité des données, ceci ne tient pas compte du contexte spécifique de notre étude.

Tableau 8 : Les 4 dimensions les plus citées dans la littérature

Dimensions	Fréquences
Completeness	19
Timeliness	18
Consistency	16
Accuracy	16

Nous avons donc repris les 22 dimensions identifiées pour les soumettre au mandant de cette étude afin de vérifier leur adéquation avec le contexte de notre étude.

5.1.3 Les dimensions identifiées par le mandant

La passation du questionnaire par notre mandant révèle une correspondance presque parfaite entre les 4 premières dimensions de la littérature et celles qu'il identifie comme les plus importantes selon son expertise (tableau 9). De facto, de fait de leur adéquation au niveau de l'importance pour la littérature et pour le mandat de cette étude, celles-ci

sont automatiquement sélectionnées. De plus, au vu de son importance dans la littérature, mais aussi qualifiée de « hautement importante » dans le questionnaire par notre participant, nous avons décidé d'intégrer la dimension *consistency* dans les dimensions obligatoirement évaluées.

Tableau 9 : Adéquation entre les dimensions de la littérature et le mandant

Dimensions	Classement	Adéquation
Accessibility	1	
Completeness	2	X
Accuracy	3	X
Timeliness	4	X

La possibilité offerte par ce questionnaire de pouvoir croiser l'état actuel des données (satisfaction) avec les besoins et les attentes (importance) de notre mandat s'avère particulièrement pertinente afin de sélectionner les dimensions utiles à notre étude (Annexe 15). En effet, en choisissant d'un côté les dimensions avec les statuts « pas du tout satisfait » et « peu satisfait » et d'un autre côté le statut « hautement important » et « très important », il est possible de mettre à jour les dimensions que nous devons utiliser dans le contexte de notre travail car elles nécessitent d'être améliorées. De plus, nous éliminons non seulement les dimensions dont la satisfaction n'est pas bonne avec une importance faible (p. ex. *concise, objectivity*), mais également les dimensions dont la satisfaction est déjà suffisante, et ceci peu importe le niveau d'importance accordé (p. ex. *informative value*) (Annexe 15).

5.1.4 Une réflexion nécessaire avant le choix final des dimensions

Avant de pouvoir effectivement répondre à notre question de recherche sur l'identification des métriques pertinentes, il est nécessaire de s'arrêter quelques instants sur un point qui a émergé pendant nos investigations.

Dans un premier temps, les articles analysés dans le chapitre concernant les DQA (c. f. 3.4.1) soulignent que l'avènement du Web des données, et que certaines de ses nouvelles dimensions, peuvent être évaluées en amont dans le processus d'évaluation et non uniquement lorsque les données sont disponibles après le scraping et le traitement. Rappelons par exemple que la sélection des sources s'est opérée pratiquement uniquement sur le critère d'accessibilité aux données (*Accessibility*) et que

les sources dont aucune donnée n'était disponible ont été écartées par notre collègue. Quant à l'actualisation des données (*Timeliness*), elle influencera directement la perception de l'actualisation ou la fraicheur des données par rapport à la tâche à accomplir dans le contexte qui est le sien. Le choix de la fréquence du scraping, plusieurs fois par jour ou uniquement une fois, aura de facto un impact sur cette dimension. Dans l'optique d'enrichir ou de compléter les données déjà à disposition, la crédibilité de la source (*Credibility*) joue également un rôle important. C'est en effet principalement sur cette dimension qu'une source sera sélectionnée. Ce constat nous pousse donc à ne pas évaluer les dimensions citées ici, car elles servent à déterminer les sources qui vont être utilisées plutôt que de servir à l'évaluation des données ainsi que de la fraîcheur des informations dont elle recèle (*Timeliness*).

En deuxième lieu, les dimensions telles que l'interprétabilité[16] (*Interpretability*), la facilité d'utilisation (*Ease of use*), l'absence d'erreurs (*Free of error*), et les métadonnées (*Metadata*), sont des dimensions qui dépendent inéluctablement du travail fait à partir du scraping. En ce qui concerne les trois premières, la mise en forme des données permet non seulement de les rendre intelligibles, mais également de révéler leur potentiel d'utilisation tout en détectant déjà des erreurs pour lesquelles une uniformisation/normalisation est possible. Quant aux métadonnées, elles viennent par exemple nous renseigner sur ce qui a été fait pendant le scraping, le traitement et sur ce qui a été mis en évidence pendant le DQA (p. ex. taux de complétude). Les informations qu'elles contiennent peuvent influencer d'autres dimensions (p. ex. *Credibility*), mais elles ne sont pas à proprement parler une dimension évaluable lors de notre DQA. Sans parler du fait que ces dimensions sont, à notre avis, hautement subjectives et que leur appréciation dépendra pour beaucoup des utilisateurs des données et de leur expertise. À ce titre, nous pensons donc que ces dimensions doivent recevoir une attention particulière lors de l'étape du traitement des données mais qu'il est difficile de les inclure dans un processus de DQA (Annexe 17).

5.1.4.1 Le traitement des données influence la qualité des dimensions

Le travail effectué par notre collègue de recherche par le biais du scraping et le traitement de certaines données par la suite est sans conteste une étape qui influence la qualité des données et l'évaluation finale de cette dernière (Druey, 2020). De ce fait, nous considérons que certaines dimensions ont déjà subi une intervention qui est censée améliorer leur qualité. On peut donc affirmer que certaines dimensions sont « raffinées » alors que d'autres restent « brutes », car elles n'ont pas été modifiées.

[16] Notre collègue de recherche s'est par exemple appliqué à traduire les informations afin de les uniformiser

Illustrons nos propos avec deux exemples : Premier cas avec l'unicité (*Uniqueness*), notre collègue a directement supprimé les entités présentes à double dans la base de données (Druey 2020). Deuxième cas pour appuyer nos propos, la validité (*Validity*) peut également être influencée lors de la phase de scraping ou de traitement. C'est le cas avec les valeurs concernant l'ETA et l'ETD ou encore le pavillon, pour lesquels notre collègue a déterminé les valeurs justes, logiques et/ou admissibles. Ce constat sur les dimensions « raffinées » s'applique à plusieurs d'entre elles que l'on retrouve dans le tableau 10.

De facto, l'intervention de notre collègue a eu un impact sur la qualité des données et modifie donc le résultat qui en découle. Dès lors, notre travail d'évaluation sur certaines dimensions peut s'apparenter à un contrôle en lien avec l'étape de traitement des données et nous oblige à considérer cette dernière comme partie intégrante d'une quelconque méthodologie d'évaluation de la qualité des données. De plus, il convient de préciser que pour l'évaluation de certaines dimensions, nous ne sommes pas en mesure de procéder à l'intégralité de l'évaluation. Dans la mesure ou certaines valeurs, pour l'exactitude (*Accuracy*) ou la cohérence (*Consistency*), doivent se mesurer en partie directement par rapport à la source ou à un référentiel (c. f. Annexe 11).

Tableau 10 : Dimensions résiduelles à évaluer suite au scraping et traitement

Dimensions	Impactées par le traitement	Données « brutes »
Completeness		X
Consistency	X	
Accuracy	X	
Uniqueness	X	
Validity	X	
Value-added		X
Precision		X

5.1.5 Les métriques pertinentes

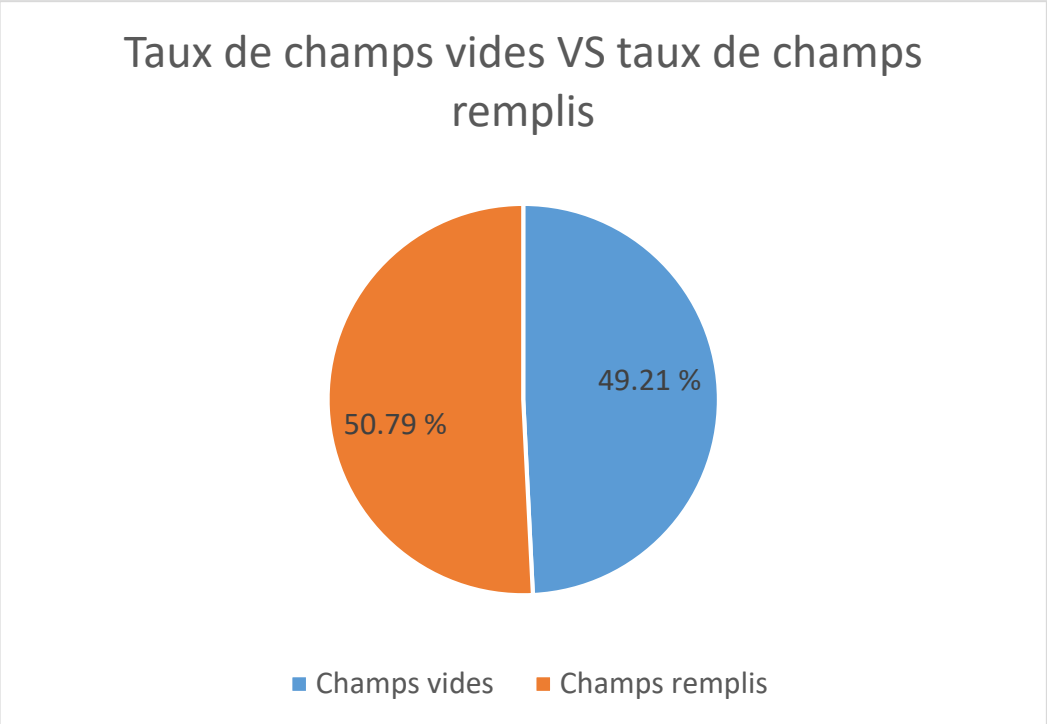
Il découle des différents constats établis dans le chapitre précédent que les dimensions suivantes, et donc les métriques qui s'y rattachent, sont pertinentes pour l'évaluation de la qualité des données en lien avec le TMMP.

5.1.5.1 Les dimensions « brutes »

5.1.5.1.1 La complétude (Completeness)

Par le biais de la complétude, il s'agit de mesurer le taux de remplissage de la base de données (Annexe 18). Ceci permet de mettre en évidence les lacunes en ce qui concerne le remplissage dont souffre la base de données que nous avons à disposition (Figure 10). Avec un taux juste supérieur à la moitié, 50,79 %, nous pouvons immédiatement constater que notre base de données souffre d'un manque évident de complétude. Ce premier constat est d'importance puisque le nombre de champs incomplets limite automatiquement toutes les dimensions que nous allons calculer par la suite. En effet, la complétude s'avère être l'une des dimensions cardinales à toute base de données : si les champs sont incomplets, il en résulte un manque d'informations qui influencent la portée et la pertinence de l'évaluation des autres dimensions.

Figure 10 : Taux de remplissage de la base de données



Ceci est particulièrement visible en faisant la distinction entre les données d'identification (Figure 11) et celles en lien avec les données de voyages (Figure 12). Avec un taux de remplissage avoisinant 68,55 % pour les données d'identification, contre 43.68 % pour les données en lien avec le voyage (Annexe 19), nous constatons une grande disparité

dans la complétude des données entre les deux catégories. Sachant que l'obtention de données en lien avec le voyage est l'une des raisons principales du projet dont fait partie notre étude, il convient d'ores et déjà de constater que la portée des informations récoltées sur les sites des ports commerciaux s'avère être limitée.

Figure 11 : Taux de complétude des données d'identification

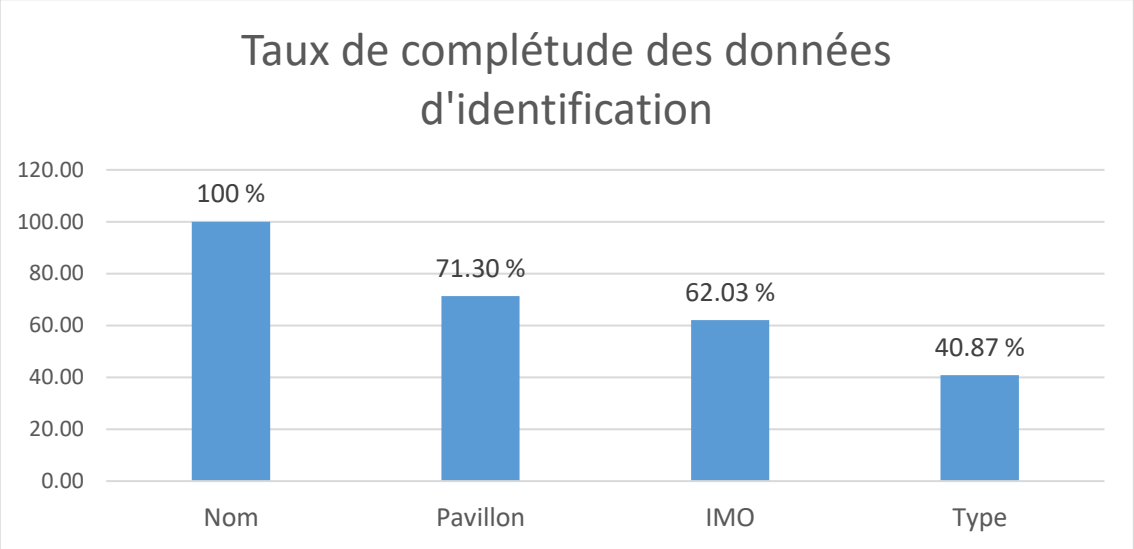
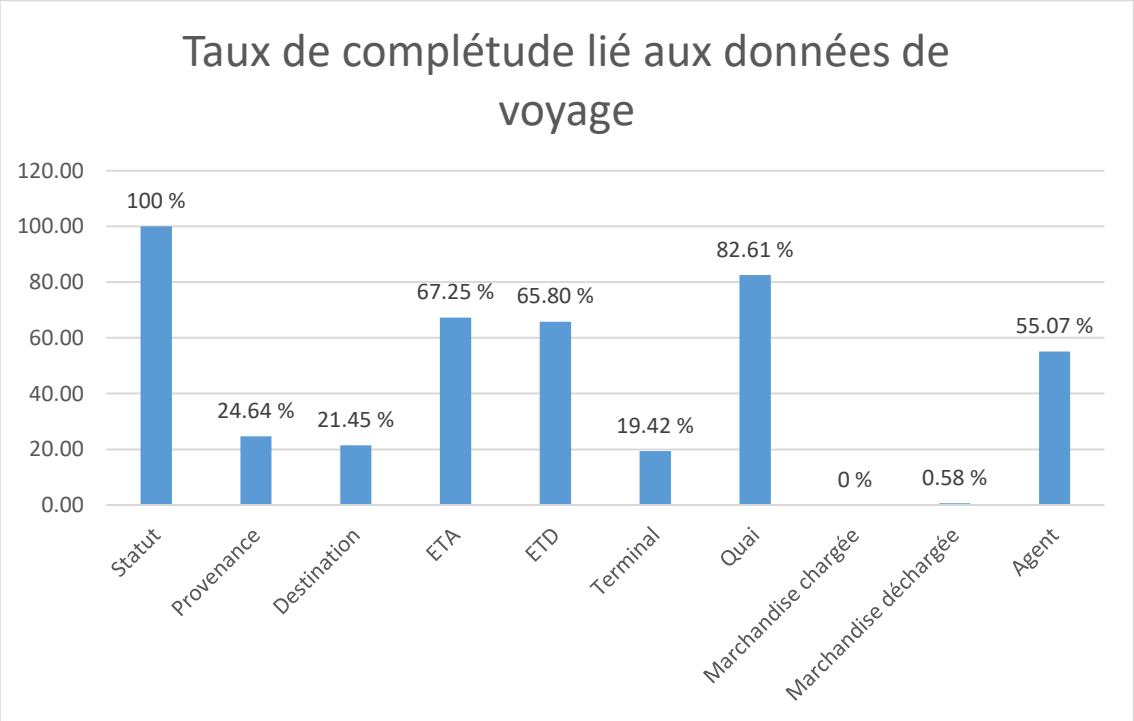


Figure 12 : Taux de complétude des données en lien avec le voyage



5.1.5.1.2 La granularité (Precision)

Nous avons défini la granularité comme le niveau de détail que contiennent les données issues de notre base de données. Ainsi nous avons identifié plusieurs attributs qui peuvent être évalués sous cet angle (tableau 11) :

- Type : le navire est précisément décrit (Tanker →Chemical Tanker)
- Provenance et destination : l'agrégation de ces deux attributs précise le trajet du navire
- ETA et ETD : ces informations peuvent être précises à la minute
- Terminal et quai : l'agrégation de ces deux attributs précise exactement le lieu d'accostage du navire
- Marchandise déchargée : il est possible de connaître la cargaison exacte du navire

De plus, nous avons décidé d'agréger certains attributs comme la provenance et la destination ainsi que le terminal et le quai. En effet, la connaissance qui peut être retirée des deux attributs ensemble s'avère être plus précise dans le cadre du voyage d'un navire que si un seul attribut était connu. Ceci est particulièrement vrai pour le couple terminal/quai : le terminal seul est d'ores et déjà une information supplémentaire, mais connaître exactement le quai apporte plus de profondeur à la connaissance que nous pouvons retirer de la base de données.

Tableau 11 : Résultats pour la précision

Variable	Formule[17]	Granularité (Précision)
Type	(141-79)/141*100	43,97
ETA	(232-0)/232*100	100
ETD	(227-0)/227*100	100
Marchandise déchargée	(2-0)/2*100	100
Terminal/Quai	(285-0)/285*100	100
Provenance/Destination	(88-17) 88*100	80,68

5.1.5.1.3 Value-added (Valeur ajoutée)

Par le biais de cette dimension, nous voulons mettre en évidence les données jusqu'à ce jour non disponibles, inconnues ou inexploitées. Mais nous nous intéressons également aux données servant à contrôler ou compléter les données existantes. Pour cette raison la dimension de la valeur ajoutée a été divisée en deux « sous dimensions » ;

[17] La formule appliquée pour l'ensemble des dimensions dans ce travail : nombre de données correctes (nombre de données vérifiées – le nombre de données fausses) / par le nombre de données vérifiées * 100.

la valeur ajoutée de contrôle et la valeur ajoutée effective. Nous avons rassemblé sous la première les attributs déjà présents dans les messages AIS (*Nom, IMO, Type, Destination, ETA*) et dans la deuxième, les nouveaux attributs issus du scraping des données (*Statut, Pavillon, Provenance, ETD, Terminal, Quai, Marchandise chargée et déchargée, Agent*) qui sont par conséquent de nouvelles données fournissant de nouvelles informations. Le calcul consiste ensuite à déterminer le ratio de champs remplis parmi ces attributs comme nous l'avons déjà fait pour la complétude. Ainsi, l'impact de la valeur ajoutée est double : elle permet de confronter, ou compléter, les données déjà connues avec les données scrapées mais aussi, d'effectivement mettre en évidence les données jusqu'alors inconnues. Nous tenons à préciser que nous développons tout ceci plus tard dans cette étude à propos de l'impact des données scrapées sur les données AIS (chapitre 5.2).

5.1.5.2 Les dimensions « raffinées »

5.1.5.2.1 Accuracy (Exactitude)

L'exactitude (*Accuracy*) vise à évaluer que l'information contenue dans les données corresponde à la réalité ou à la situation du monde réel par rapport à une référence connue (Annexe 11). Par exemple, que l'IMO d'un navire soit bien un numéro ou alors que le type d'un navire soit bien un type de navire, etc. Nous avons appliqué ce principe aux attributs ou il était possible de contrôler l'information que nous avons à disposition à savoir *Pavillon, IMO, Type de navire, statut, ETA, ETD, Marchandise chargée et déchargée* (Annexe 20). En revanche, nous avons été dans l'impossibilité de faire ce contrôle pour les attributs *Nom, Provenance, Destination, Terminal, Quai et Agent*. Car dans ces cas, nous ne pouvons pas savoir si les valeurs dans notre base de données représentent effectivement la réalité ou non. A souligner que les résultats parfaits des champs investigués pour cette dimension s'expliquent par l'intervention et l'uniformisation des données par notre camarade de recherche (Annexe 21).

5.1.5.2.2 Consistency (Cohérence)

La cohérence se caractérise par le fait qu'elle permet non seulement de comparer les valeurs contenues dans une base de données par rapport à une autre base de données, mais surtout de vérifier la non-contradiction des valeurs au sein d'une même base de données. Ainsi, nous pouvons vérifier que certaines valeurs à notre disposition dans cette dernière ne soient pas contradictoires entre elles. Cependant, au vu du manque de complétude de notre base de données, nous ne pourrons pas analyser toutes les relations possibles entre les attributs. Néanmoins, nous nous sommes attachés à contrôler les possibles contradictions entre les attributs suivants :

Tableau 12 : Vérification de l'attribut statut

Vérifications	Contradictions
Statut parti vs ETD[18]	22
Statut partant vs ETD[19]	22
Statut A quai/arrivé vs ETA/ETD[20]	2
Statut attendu/notifié vs ETA[21]	1

Tableau 13 : Résultats pour la cohérence

Variable	Formule	Cohérence (Consistency)
Statut (tous)	(338[22]-47)/338*100	86,10
ETD[23]	(119-0)/119*100	100
ETA[24]	(119-0)/119*100	100
Marchandise déchargée[25]	(2-0)/2*100	100

Les résultats présentés dans le tableau 12 mettent en évidence l'importance de procéder à la vérification des relations entre les données. Car lorsque que nous comparons le statut des navires avec les autres données à disposition dans la base de données, nous constatons des incohérences. Par exemple, le jour du scraping, plus d'une vingtaine de navire ont le statut « parti » alors que l'attribut *Estimated time of departure* est postérieur à cette dernière. De plus, une vingtaine de navires ont également le statut « partant » pendant plusieurs jours. A ce stade de notre étude, nous ne pouvons qu'émettre

[18] Le jour du scraping, un navire ne devrait pas avoir le statut « Parti » si l'ETD est dans le futur
[19] Un navire ne peut avoir un statut partant pendant plusieurs jours
[20] Le statut à quai/arrivé est confirmé soit par l'ETA ou ensuite par l'ETD
[21] Le statut est confirmé par l'ETA
[22] Du total de 345 navires, les données suivantes ont été soustraites : 5 navires n'ont pas de données ETA ou ETA et 2 ont un statut qui ne peut pas être vérifié
[23] Un navire ne peut quitter un port avant d'y être arrivé
[24] Un navire ne peut arriver au port après son ETD
[25] La marchandise déchargée est en adéquation avec le type de navire qui la transporte

l'hypothèse que ces incohérences sont probablement dues à l'absence de mise à jour des données sur les sites des ports sélectionnés. Cette explication nous semble à privilégier dans le cas « Statut parti vs ETD » ou il se peut que les navires soient bel et bien partis avant ce qui avait été estimé sans que les données aient été mise à jour. En revanche, le statut « partant vs ETD » nous paraît plus difficile à comprendre. Ce statut pourrait tout aussi bien s'expliquer par un quelconque retard qui empêche un navire d'appareiller.

Sans toutes les énoncer ici, d'autres relations auraient pu être examinées par exemple, que le quai annoncé soit effectivement dans le terminal décrit qui lui-même est bien dans le port identifié (Quai →Terminal →Port). Ou encore, que le nom du navire corresponde bel et bien au pavillon et au numéro IMO (Nom →Pavillon →IMO). Mais le manque d'information (cf. complétude) ainsi que de l'absence d'un référentiel pour les ports de cette étude nous empêche de pousser nos investigations plus loin que les relations que nous avons mises en évidence.

5.1.5.2.3 Uniqueness (Unicité)

L'unicité de la base de données (*Uniqueness*) est utile pour déterminer le nombre de doublons présents dans cette dernière. Comme nous l'évoquions auparavant, les entrées correspondantes aux mêmes navires et avec les mêmes informations ont déjà bénéficié d'un traitement lors de la création de la base de données (Druey 2020). Par conséquent, la mise au jour de 35 navires présents deux ou trois fois dans notre échantillon peut s'expliquer par exemple de la manière suivante :

- Il y a un changement de statut (p. ex. A quai →parti)
- S'il n'y a pas de changement de statut, mais que nous avons deux dates de scraping différentes, une information précédemment manquante a été mise à jour (p. ex. ETA, ETD, Terminal, Quai, etc.)
- Un navire peut naviguer entre deux ports sélectionnés dans cette étude
- La fonction du bateau explique sa présence à plusieurs reprises dans la base de données, ce qui est par exemple le cas des remorqueurs

En appliquant les critères que nous venons de mentionner, l'échantillon que nous avons utilisé dans notre recherche ne rend compte d'aucun doublon, soit un taux d'unicité de 100 % (tableau 14).

Tableau 14 : Résultat pour l'unicité

Variable	Formule	Unicité (Uniqueness)
Toutes les entités (navires)	(345-0)/345*100	100

5.1.5.2.4 Validity (Validité)

Par le biais de la validité, nous contrôlons que les valeurs de la base de données respectent une syntaxe ou un format prédéfini comme le n° IMO, qui doit se composer de 7 chiffres, ni plus ni moins. Nous avons aussi pu examiner tous les attributs sauf les marchandises chargées pour lesquelles nous n'avons aucune information. De plus, à l'image de certaines autres, l'évaluation de cette dimension est grandement influencée par le traitement effectué par notre camarade de recherche (Annexe 20). Dans le cas qui nous intéresse, les fautes constatées proviennent d'une part d'un statut contradictoire (« Attendu/à quai ») et d'autre part du champ nom contenant également le type du navire (e. g Meteor pousseur).

5.1.6 Synthèse et remarques intermédiaires

5.1.6.1 Le problème de la complétude

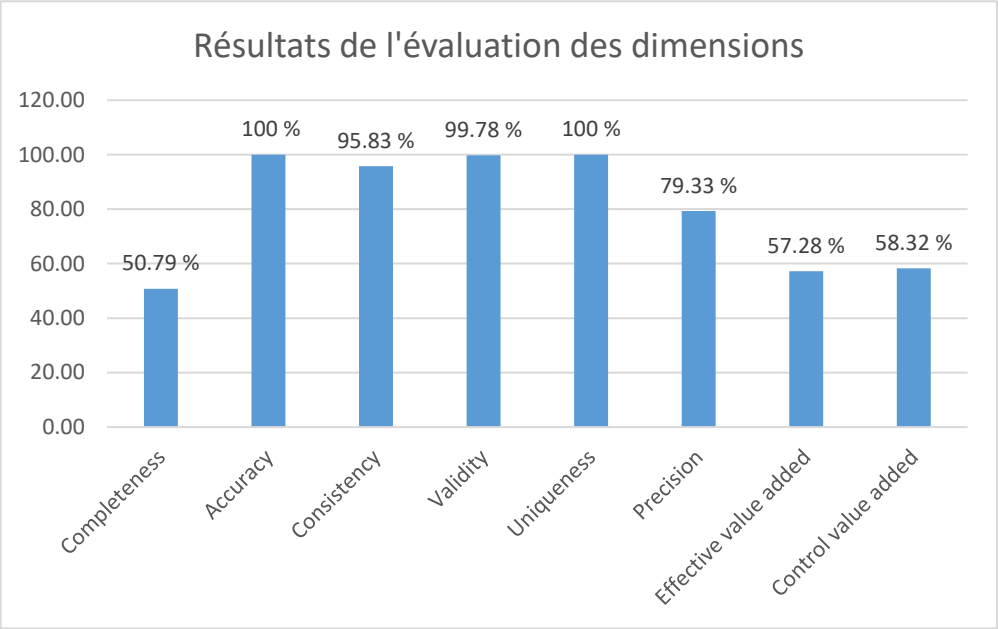
Nous l'avons déjà évoqué, le faible taux de complétude est l'élément principal que nous pouvons mettre en évidence par le biais de notre analyse. Cette dimension influence sans aucun doute possible les autres dimensions comme la cohérence (*Consistency*). En effet, à cause de cela, nous n'avons pas pu examiner certaines relations au sein de la base de données alors que cela aurait pu être possible avec davantage de données. En outre, des attributs tels que la provenance, la destination ou encore ceux qui se rapportent aux mouvements de marchandise souffrent d'un manque évident de valeurs qui n'est pas sans conséquence sur le reste de notre enquête. Par conséquent, et en reprenant l'exemple des mouvements des marchandises, il est nécessaire de nuancer, ou du moins de garder en tête que la représentativité des données est fortement limitée dans la base de données issues du scraping (une trentaine de valeurs) et dans notre échantillon (deux valeurs).

5.1.6.2 Le traitement et ses conséquences

La normalisation/l'uniformisation, ou ce que nous avons identifié comme le traitement des données, réalisé par notre collègue soulève également quelques questions. Cette étape antérieure à notre analyse s'avère être d'une certaine importance. En effet, celle-ci permet déjà de remédier à plusieurs lacunes dans les données par rapport à certaines dimensions comme pour l'exactitude, la validité ou encore l'élimination des doublons (Figure 13). Dans tous les cas, ce traitement augmente grandement la qualité des données à disposition puisque ces dimensions sont pratiquement toutes à 100 %. Cependant, dans le cas de l'exactitude (*Accuracy*) et de la consistance (*Consistency*), il convient que ce traitement examine attentivement que les données récoltées soient celles réellement exprimées à l'origine car ceci est l'un des prérequis pour s'assurer de la qualité d'une base de données (Annexe 11). De plus, la création d'un référentiel

normalisé afin de traiter les données issues des sources selon des critères établis est primordiale. En effet, ceci permet non seulement de limiter une surabondance de valeurs possibles dans les champs, mais aussi repérer les valeurs qui nécessiteront un travail d'uniformisation (Druey 2020).

Figure 13 : Résultats de l'évaluation des dimensions



5.1.6.3 L'analyse de la cohérence interne de la base de données

La cohérence se caractérise tant par le fait que les données récoltées représentent effectivement les données d'origine ainsi que la non-contradiction de celles-ci entre elles à l'intérieur de la base de données. Nous nous sommes employées à mettre ce dernier point en évidence dans l'évaluation de la dimension de la cohérence. Mais pour ce faire, il est nécessaire d'imaginer les règles potentielles dans la base de données qui puissent être un indice de contradiction des données. Dans cette recherche, nous avons par exemple mis au jour qu'un bateau ne peut décemment pas partir d'un port sans y être arrivé, etc. La complétude des données que nous avons à disposition, a clairement déterminé les relations que nous avons pu examiner, mais une base encore plus complète pourrait certainement permettre d'identifier d'autres contradictions. Avec cette dimension de la cohérence, il est véritablement important de réfléchir aux relations entre les données et d'analyser les données à disposition sous cet angle. Par conséquent, la fixation de normes internes à la base de données permet de vérifier la véracité des informations et augmente la confiance que l'on peut y accorder.

5.1.7 Mise en application du modèle TMMP pour l'indice de qualité

Après avoir effectué tous les calculs nécessaires, et ceci pour toutes les dimensions contextuelles retenues dans notre modèle grâce à notre questionnaire, il est ensuite

possible de procéder à l'évaluation intermédiaire pour chacune de celles-ci (Annexe 21). Cependant, avant de procéder à ce calcul, il est nécessaire d'attribuer des poids aux attributs (colonnes). Dans le cadre de notre recherche, nous avons choisi ces poids en fonction d'une part de l'importance d'un attribut dans la dimension à évaluer et d'autre part des valeurs disponibles pour l'évaluation (Annexe 21).

Pour la deuxième phase du calcul, visant à finalement connaître l'indice de qualité globale de la base de données, les pondérations par dimension ont été déterminées selon le classement établi par notre mandant et selon ses réponses à notre questionnaire (tableau 15).

Tableau 15 : Résultat final du modèle TMMP

	Résultat par dimension	**Classement dans le questionnaire**	**Pondération**	**Résultat final**
Completness[26]	**50.79**	2	0,30	15.24
Accuracy	100	3	0,20	20
Uniqueness[27]	**100**	5	0,15	15
Validity	99,78	8	0,15	14,97
Effective value added	57,27	11	0,05	2.86
Control value added	58,32		0,05	2,92
Consistency	95,83	12	0,05	4,79
Precision	79,33	17	0,05	3,97
Total			**1**	**<u>79.74 %</u>**

Ainsi, selon toutes les étapes et les calculs décrits au préalable, nous constatons que la base de données créée pour les besoins de cette recherche, et donc les données issues

[26] Un poids identique a été appliqué à la complétude, une moyenne est donc suffisante
[27] L'unicité se calcule sur l'ensemble des entités (lignes) de la base de données

du scraping des sites internet des ports commerciaux sélectionnés pour cette étude, ont un indice de qualité globale de 79,74 %.

5.1.8 Synthèse des résultats finaux

Notre méthode nous a permis de sélectionner les dimensions, et donc les métriques pertinentes, en fonction non seulement du contexte du TMMP, mais aussi au regard des besoins et des attentes de notre mandant. Les sept dimensions choisies nous permettent ensuite de procéder au calcul d'un indice global de la qualité des données issues du web scraping de notre collègue chercheur. Cependant, la tâche effectuée soulève trois points d'attention particulièrement importants.

5.1.8.1 Benchmark(s)

Nous avons démontré qu'il était possible d'objectiver la qualité d'une base de données grâce aux dimensions de la qualité et leurs métriques en se basant sur les besoins de l'utilisateur des données. Pourtant, bien que le résultat soit intelligible, sa portée s'appréhende difficilement car nous ne possédons pas d'autres métriques, indice ou benchmark pour le confronter et le comparer.

Dans le cadre de nos investigations pour cette étude et parmi les articles scientifiques sur lesquels elle repose, aucune proposition de benchmark en lien avec l'évaluation individuelle des dimensions n'est présentée et le même constat s'applique en ce qui concerne la phase finale de l'évaluation. Pourtant, comme nous l'évoquions auparavant dans ce travail, l'appréhension de la qualité dans un processus de DQA passe irrémédiablement par la comparaison et la confrontation des résultats à des valeurs ou des seuils. Il est clair que le caractère exploratoire et itératif de notre recherche vise également à souligner et mettre en évidence d'éventuelles lacunes quant au champ d'études. Nous pensons dès lors qu'un indice de référence pour confronter les sources de données dans le contexte du TMMP est une nécessité sinon comment complétement appréhender le résultat ?

Toutefois, nous ne pensons pas que ce manquement remette en question la portée de cette recherche. Nous affirmons cependant qu'un futur benchmark doit, à l'image de cette étude et du projet dans sa globalité, prendre en compte la réalité du champ d'études et les besoins de l'utilisateur final.

5.1.8.2 Le résultat repose sur notre mandant

La détermination de la pondération finale des dimensions repose uniquement sur les besoins de notre utilisateur selon ses réponses fournies à notre questionnaire. Son rôle d'expert dans le domaine nous pousse bien sûr à accorder du crédit à ces dernières, mais il subsiste toutefois l'idée que la pondération pourrait être différente avec

l'interrogation d'un panel d'experts et d'utilisateurs plus large. Car nous relevons certaines contradictions dans les réponses obtenues et notamment par rapport à la classification des dimensions. En effet, notre participant a évalué des dimensions comme « peu importantes » alors qu'il les a très bien classées au moment d'établir le classement des dimensions les plus importantes dans son domaine d'activité (Annexe 15). Nous pensons que ce genre de biais pourrait être évité grâce à la passation du questionnaire par plus d'individus et que cela serait plus représentatif pour déterminer non seulement les dimensions en lien avec le TMMP, mais aussi dans le choix des pondérations.

5.1.8.3 Pondérations pour tester le modèle

Pour tester notre modèle de calcul et de pondérations TMMP, nous avons attribué des poids à l'évaluation des dimensions qui sont avant tout pragmatiques. En effet, l'évaluation individuelle de ces dernières a été logiquement influencée par les données à notre disposition, et principalement par cet état de fait. Par exemple, aucun poids n'a été attribué à la complétude car nous avons considéré que les attributs sont d'égale importance. En revanche, pour la cohérence, toute l'évaluation repose sur les attributs que sont le *statut, l'ETA, ETD* alors que la *marchandise* n'apporte aucune connaissance particulière. Il est donc de bon sens d'attribuer des poids bien plus élevés à ces statuts alors que le dernier ne nous apporte aucune information.

Cependant, l'attribution des poids n'est pas une tâche aisée sans expertise du contexte d'utilisation des données et sans avoir clairement défini les attributs qui ont une valeur plus conséquente que d'autres. Tout ceci, sans oublier que les attributs sélectionnés peuvent grandement souffrir d'un manque de valeurs dans les champs de la base de données et donc voir leur importance ainsi que leur pertinence grandement diminuer. Nous pensons donc que l'attribution de pondération dans notre modèle doit se faire de manière plus raisonnée par exemple grâce à la formulation des besoins précis de l'utilisateur des données.

5.2 Question de recherche n 2 : l'impact du Web scraping sur la qualité des données

- Est-ce que les données récoltées par le biais d'outils de Web Mining auprès des sites internet de certains ports permettent d'améliorer la qualité des données ?

Les données actuellement à disposition de notre mandant sont les données AIS et les données de navigation calculées en fonction des mouvements des navires. Dès lors, l'utilité du scraping ne vise pas uniquement à récolter des données inconnues ou manquantes, mais également de permettre la vérification des données déjà connues.

Bien que les données de voyages soient au cœur de ce projet de recherche et des besoins de notre mandant, les données d'identification offrent l'opportunité de vérifier les données déjà à disposition. Celles-ci ont donc également une valeur. En effet, comme nous l'avons démontré au début de cette étude, la fiabilité des données fournies par le système AIS est sujette à caution. Bien que nous ne connaissions pas exactement l'ampleur des lacunes dans les données de notre mandat, nous sommes partis du postulat que la dimension de la valeur ajoutée (*value-added*) pouvait donc être tantôt l'indicateur de l'apport de nouvelles données aux données à disposition, mais également une variable de contrôle des données en possession de notre mandant. Par conséquent, nous avons décidé de scinder cette dimension en deux éléments : la valeur ajoutée effective et la valeur ajoutée de contrôle.

5.2.1 La valeur ajoutée de contrôle

La valeur ajoutée de contrôle doit se comprendre comme le moyen de vérifier ou de compléter les données AIS en la possession de notre mandant. En effet, les attributs *Nom, IMO, Type, Destination et ETA* font partie des informations contenues dans les messages AIS (Annexe 2). Par conséquent, elles ne sont à proprement parlé pas des nouvelles valeurs, mais des données qui peuvent potentiellement compléter des attributs existants. Leur impact réside donc dans l'opportunité qu'elles offrent d'effectuer une vérification des données connues, de compléter des champs inconnus ou des valeurs manquantes dans les messages AIS. Avec une valeur de 58.32 %, les données scrapées peuvent s'avérer utiles dans le but de contrôler ou de compléter les données existantes (Annexe 21). Cependant, nous n'avons pas pu effectuer la comparaison dans la cadre cette recherche. En effet, pour ce faire nous aurions dû posséder les données AIS des navires effectivement sélectionnés dans notre échantillon afin de pouvoir procéder à la comparaison et l'évaluation des données.

5.2.2 La valeur ajoutée effective

La valeur ajoutée effective quant à elle se comprend comme l'apport effectif de données jusqu'alors non disponibles ou inconnues dans les données à disposition de Riverlake. De sorte que les attributs *Statut, Pavillon, Provenance, ETD, Terminal, Quai, Marchandises chargées et déchargées* ainsi que *l'agent* ne figurent pas dans les messages AIS (Figure 14). Ces derniers peuvent donc être considérés comme de nouveaux attributs dont les valeurs représentent de nouvelles données. Cet apport est comparable à la dimension de la complétude. En effet, une fois les nouveaux attributs identifiés comme à « valeur ajoutée », le nombre de valeurs présentes dans les champs de ces attributs sont de facto les données nouvellement exploitables. Dans le cadre de cette étude, la valeur ajoutée effective est de 57.28 % selon les pondérations accordées

aux attributs (Annexe 21). Par exemple, un poids plus important est accordé aux attributs représentant une plus-value particulièrement appréciable tels que l'*ETD*, le *Terminal* et le *Quai*. Alors que dans le cas contraire, des attributs comme le pavillon sont moins bien considérés.

5.2.3 Valeur ajoutée finale

Le choix de partager la dimension de la valeur ajoutée a pour conséquence de légèrement modifier le résultat de l'indice global de qualité que nous avons imaginé. En effet, puisque nous séparons cette dernière en deux « sous dimensions » et nous y attribuons des pondérations différentes, le résultat passe de 79,04 % à 79,74 %. Cependant, la pondération finale de la dimension valeur ajoutée déterminée par le biais de notre questionnaire respecte toujours les besoins de notre mandant. Nous avons tout simplement divisé cette dernière entre les deux « sous-dimensions » soit un poids de 0.05. De la sorte, nous postulons que cette dimension de la valeur ajoutée est plus représentative du contexte dont elle est issue et permet de mieux appréhender l'impact des données récoltées sur le Web par rapport aux données déjà à disposition.

5.2.4 Un impact positif

5.2.4.1 La valeur ajoutée

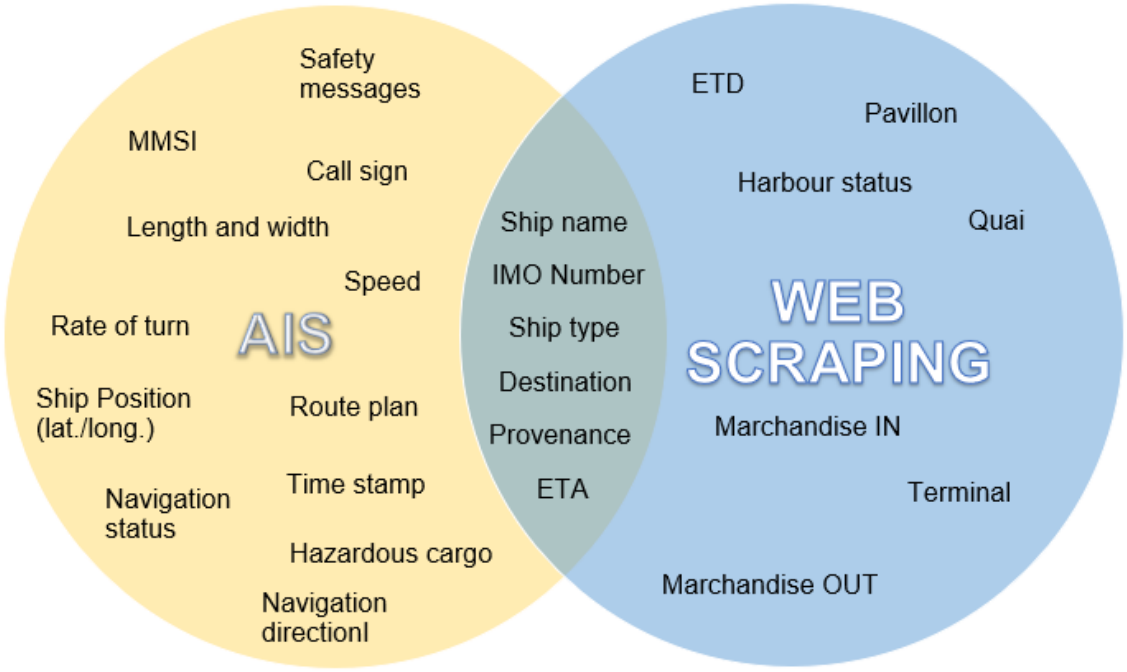
Avec ce que nous venons d'exposer, nous avons démontré que les données issues des ports sélectionnés pour ce projet de recherche ont un impact positif sur la qualité des données. Si tel n'était pas le cas, les deux sous-dimensions de la valeur ajoutée seraient proches de zéro soulignant ainsi un apport très limité. Dans le cadre de notre recherche, les deux taux s'approchent des 60 %. Pourtant, à l'image de ce que nous avons déjà évoqué dans le chapitre 5.1, ce résultat souffre également de la possibilité d'être comparé à d'autres résultats. Même si une interprétation rapide de ces résultats laisse entrevoir que les données scrapées sont prometteuses, il s'agit toutefois d'une évaluation qui n'est que partiellement objective et qui repose sur un scraping unique. Un jugement plus approfondi serait néanmoins possible lorsque le présent projet aura procédé à d'autres scrapings et d'autres évaluations de base de données. Ainsi, l'expérience accumulée au niveau de l'utilisation du modèle, mais également son application sur une plus grande quantité de données devrait permettre d'établir des métriques afin d'en vérifier la qualité.

5.2.4.2 De nouvelles données pour de nouvelles connaissances

Le travail effectué sur la base de données issues des ports commerciaux choisis pour ce projet de recherche met en évidence des données qui étaient inconnues par le biais des données AIS (Figure 14). Les informations glanées permettent dorénavant d'avoir une meilleure connaissance du voyage des navires, mais aussi de récolter des données

très utiles quant à l'accostage des navires. En effet, il est par exemple tout à fait concevable que les attributs *Type, Terminal et Quai* aident à déduire la marchandise que transporte un navire grâce aux infrastructures de chargement ou de déchargement présentent dans le port et sur le quai. De sorte que les besoins pour le déchargement et le chargement d'un tanker ne soient pas les mêmes que pour les opérations liées à un vraquier. En outre, il est imaginable qu'un agent soit spécialisé dans le transport d'un seul type de marchandise, laissant entrevoir d'importantes opportunités quant aux données scrapées.

Figure 14 : Apport du Web Scraping aux données AIS



6. Recommandations

6.1 La nécessité des benchmarks

Par le biais de cette étude, nous avons montré que le modèle contextuel développé fonctionne selon les attentes de notre mandant. Toutefois, afin de le rendre tout à fait implémentable, il est nécessaire de réfléchir à la création de benchmarks pour l'évaluation individuelle des dimensions ainsi que pour l'indice global de la qualité.

Ceci pourrait par exemple être réalisé rétroactivement grâce à la conduite de nouvelles évaluations permettant ainsi de refléter les spécificités du contexte du TMMP tout en se basant sur une plus grande quantité de données. Car il ne faut pas oublier que de nombreux facteurs influent sur le transport maritime comme par exemple les évènements météorologiques, macroéconomique, politique, etc. Mais avec la profusion

de données disponibles sur internet, il est possible d'utiliser des données historiques couvrant une plus longue période. En effet, rappelons que notre étude ne se concentre que sur des données récoltées en l'espace de trois jours. Alors que des benchmarks calculés à différents moments et sur de plus longues périodes, gagneraient en légitimité et en pertinence. Ainsi, il serait envisageable de créer un indice sur l'ensemble des scraping réalisé dans le but d'avoir un benchmark au plus proche de la réalité de notre champ d'étude.

De plus, le modèle développé dans notre travail n'est à l'heure actuelle qu'une proposition d'évaluation. À notre sens, l'approche exploratoire de notre étude et les conclusions que nous pouvons en tirer méritent d'être confirmées par notre mandant afin d'en vérifier l'adéquation avec ses besoins et son contexte d'affaires. Enfin, un indice global de qualité, comparé à un benchmark, représente également une manière efficace de monitorer l'évolution d'une source ou d'une base de données. Une quelconque variation dans cet indice attirerait l'attention et obligerait l'utilisateur des données à procéder par exemple à l'ajustement du scraping ou le changement d'une source de données. Ainsi, en identifiant les problèmes liés aux données, la qualité des données en lien avec le TMMP serait véritablement dans un processus d'amélioration continue comme nous le décrivons au chapitre 3.4.1.1.

6.2 Propositions de pondérations

6.2.1 Augmenter le poids de la cohérence

Le résultat de l'indice de qualité globale repose sur les pondérations déterminées en fonction des réponses fournies par notre mandat lors du questionnaire. Bien que celles-ci nous paraissent globalement pertinentes, la pondération accordée à la cohérence mérite d'être revue. En effet, lorsque nous avons évalué cette dimension avec l'aide de notre modèle, nous avons remarqué que des données récoltées peuvent être contradictoires (tableau 13). À notre sens, des données fournissant des informations contradictoires sont plus à même de remettre en question la qualité de la connaissance que l'on peut justement attendre d'une base de données que par exemple sa validité. Bien que cette dernière soit importante, la phase de traitement des données permet d'intervenir directement sur celle-ci, comme cela a été fait par notre collègue de recherche. Des données peuvent être exactes, représenter le monde qu'elles décrivent, mais être en opposition comme nous l'avons démontré dans notre analyse.

L'évaluation de la cohérence nécessite, à notre sens, plus d'attention que les autres dimensions car il faut appréhender les relations entre les données contenues dans la base de données. Ceci demande certes plus d'efforts, mais apporte beaucoup plus de

pertinence à l'information qu'il est possible d'extraire des données et de facto d'en augmenter la qualité.

6.3 Traitement

Le traitement réalisé par notre collègue est une étape primordiale car sans cette intervention, l'évaluation des données se serait avérée plus longue et plus fastidieuse. L'impact de ce dernier sur les dimensions telles que la validité ou l'unicité est clairement perceptible par le biais des résultats obtenus. Pour ainsi dire, le traitement et l'uniformisation des données permettent de se consacrer avec plus d'assiduité aux autres dimensions identifiées dans ce projet de recherche.

Cependant, nous tenons à relever que ce travail de traitement et d'uniformisation des données doit se baser sur des référentiels reconnus ou identifiés comme pertinents et fiables pour la suite du projet. Prenons par exemple le pavillon, dans l'idéal ce dernier doit reposer sur un format/syntaxe qui soit reconnu comme tel afin non seulement de faciliter le traitement, mais aussi de correspondre aux normes en vigueur.

Le traitement des données issues du scraping doit donc être appréhendé comme une étape à part entière dans l'évaluation de la qualité de ces mêmes données. En effet, il rend les données (plus) intelligibles, augmente leur qualité, donne un aperçu de l'information disponible tout en facilitant ensuite l'évaluation des données.

6.4 Évaluer la source

Bien que ce soit le résultat de l'indice de qualité globale qui révèle la qualité d'une source, nous pensons qu'un travail en amont doit être fait sur la source. Sans que celui-ci soit trop important ou trop contraignant, une source doit être examinée afin de déterminer d'où proviennent les données, si ces dernières ont été modifiées, etc. Nous retrouvons ici l'idée d'avoir des données sur les données (métadonnées). Ainsi cette prise d'information quant à la source peut permettre de grandement contribuer à diminuer par exemple le traitement des données tout en s'assurant que les données qui vont être utilisées bénéficient déjà d'un certain niveau de qualité. Nous pensons par exemple ici à la mise à jour des données sur les sites internet. En effet, connaître la fréquence à laquelle celles-ci sont changées permet d'avoir des données plus fraîches et donc plus utiles. Récolter des données qui ne soient pas de bonne qualité augmenterait à coup sûr le temps de traitement sans assurer qu'une fois cette étape terminée, les données soient exploitables. En outre, comme le souligne notre collègue de recherche, la ou les sources des données disponibles sur les sites des ports n'est dans la plupart des cas pas signalée. Il conviendrait donc de s'assurer que ces dernières ne proviennent pas

directement du système AIS car cela pourrait grandement limiter la portée des données récoltées pour cette étude (Druey 2020).

7. Conclusion

Bien que notre étude soit exploratoire et donc limitée par certaines observations faites pendant le processus de recherche et nécessite des approfondissements, notre recherche a permis de répondre complètement ou en partie aux questions suivantes :

- Quelles sont les métriques pertinentes quant à l'évaluation de la qualité des données ?
- Est-ce que les données récoltées par le biais d'outils de Web Mining auprès des sites internet de certains ports permettent d'améliorer la qualité des données ?

Dans un premier temps, avant de pouvoir effectivement identifier les métriques pertinentes à l'évaluation de la qualité, nous avons dû mener des recherches afin de mettre au jour les dimensions constitutives de la qualité des données. Pour ce faire, nous avons procédé par étape en commençant par mettre en évidence les dimensions citées dans plus de vingt articles scientifiques. Basée sur cette dernière, nous avons ensuite interrogé notre mandant pour qu'il identifie et sélectionne les dimensions les plus à même de rendre compte de son contexte de travail et de ses besoins en matière de qualité des données. De ce processus, résulte une évaluation qui se base sur les dimensions suivantes *: la complétude, l'exactitude, la cohérence, la validité, l'unicité, la précision et la valeur ajoutée*. Par le biais du modèle proposé dans cette recherche, comprenant une méthode de calcul et des pondérations pour les dimensions ainsi qu'une pondération finale prenant spécifiquement en compte le contexte de notre mandant, nous avons pu déterminer que les données récoltées auprès des onze ports commerciaux atteignent un indice de qualité globale de 79,74 %. Malheureusement, l'absence d'un côté de mesures permettant de faire une comparaison quant aux dimensions (Figure 13) et de l'autre l'absence d'un benchmark quant à l'indice global (Tableau 15) ne nous permette pas de complètement appréhender la portée du résultat de cette étude. Mais cela se comprend par le caractère intrinsèquement exploratoire de ce travail. Cette limite momentanée ne remet toutefois pas en question le modèle développé pour l'évaluation des données en lien avec le TMMP dans ce travail de recherche.

Dans un deuxième temps, la dimension de la valeur ajoutée nous permet de juger de l'impact des données scrapées sur les données AIS déjà à disposition. Divisée en deux sous-dimensions, la valeur ajoutée de contrôle pour vérifier ou compléter les données déjà connues ainsi que la valeur ajoutée effective qui concerne des données inconnues

et donc de nouvelles sources d'information. Dans les deux cas avec un taux approchant les 60 %, ces dimensions démontrent que les données issues des ports commerciaux améliorent la qualité des données AIS. De plus, les données inconnues jusqu'alors permettent d'imaginer de nouvelles relations entre les données et ainsi générer de nouvelles connaissances qui seront très utiles quant au trafic maritime de matières premières.

Bibliographie

AARSÆTHER, Karl Gunnar et MOAN, Torgeir, 2009. Estimating Navigation Patterns from AIS. In : *The Journal of Navigation*. octobre 2009. Vol. 62, n°4, p. 587-607. DOI 10.1017/S0373463309990129.

ADLAND, Roar, JIA, Haiying et STRANDENES, Siri P., 2017. Are AIS-based trade volume estimates reliable? The case of crude oil exports. In : *Maritime Policy & Management*. 4 juillet 2017. Vol. 44, n°5, p. 657-665. DOI 10.1080/03088839.2017.1309470.

ARDAGNA, Danilo, CAPPIELLO, Cinzia, SAMÁ, Walter et VITALI, Monica, 2018. Context-aware data quality assessment for big data. In : *Future Generation Computer Systems*. 1 décembre 2018. Vol. 89, p. 548-562. DOI 10.1016/j.future.2018.07.014.

ARGUEDAS, Virginia Fernandez, PALLOTTA, Giuliana et VESPE, Michele, 2018. Maritime Traffic Networks: From Historical Positioning Data to Unsupervised Maritime Traffic Monitoring. In : *IEEE Transactions on Intelligent Transportation Systems*. mars 2018. Vol. 19, n 3, p. 722-732. DOI 10.1109/TITS.2017.2699635.

AZEROUAL, Otmane et ABUOSBA, Mohammad, 2017. Improving the Data Quality in the Research Information Systems. In : *International Journal of Computer Science and Information Security*. 1 novembre 2017. Vol. 15, p. 82-86.

AZEROUAL, Otmane, SAAKE, Gunter et WASTL, Jürgen, 2018. Data measurement in research information systems: metrics for the evaluation of data quality. In : *Scientometrics*. 1 juin 2018. Vol. 115, n°3, p. 1271-1290. DOI 10.1007/s11192-018-2735-5.

BALDUZZI, Marco, PASTA, Alessandro et WILHOIT, Kyle, 2014. A security evaluation of AIS automated identification system. In : *Proceedings of the 30th Annual Computer Security Applications Conference* [en ligne]. New Orleans, Louisiana, USA : Association for Computing Machinery. 8 décembre 2014. p. 436–445. [Consulté le 25 avril 2020]. Disponible à l'adresse : https://doi.org/10.1145/2664243.2664257.

BATINI, Carlo, CAPPIELLO, Cinzia, FRANCALANCI, Chiara et MAURINO, Andrea, 2009. Methodologies for Data Quality Assessment and Improvement. In : *ACM Computing Surveys*. 1 juillet 2009. Vol. 41, n°3. DOI 10.1145/1541880.1541883.

BEHKAMAL, Behshid, KAHANI, Mohsen, BAGHERI, Ebrahim et JEREMIC, Zoran, 2014. A Metrics-Driven Approach for Quality Assessment of Linked Open Data. In : *Journal of theoretical and applied electronic commerce research*. mai 2014. Vol. 9, n° 2, p. 64-79. DOI 10.4067/S0718-18762014000200006.

BERTI-EQUILLE, Laure, 2004. (PDF) Qualité des données. In : *ResearchGate* [en ligne]. [Consulté le 25 avril 2020 b]. Disponible à l'adresse : https://www.researchgate.net/publication/220438866_Qualite_des_donnees.

CAI, Li et ZHU, Yangyong, 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. In : *Data Science Journal*. 22 mai 2015. Vol. 14, n 0, p. 2. DOI 10.5334/dsj-2015-002.

CAN, François, GAUDINAT, Arnaud, et THEODORO, Douglas, 2020. PreciseIntelligence: Big Data Analytics for Comprehensive Global Trade Flow Intelligence [Innosuisse - online application].

CAPPIELLO, Cinzia, FRANCALANCI, Chiara et PERNICI, Barbara, 2004. Data quality assessment from the user's perspective. In : *Proceedings of the 2004 international workshop on Information quality in information systems* [en ligne]. Paris, France : Association for Computing Machinery. 18 juin 2004. p. 68–73. [Consulté le 12 mai 2020]. Disponible à l'adresse : https://doi.org/10.1145/1012453.1012465.

CAZZANTI, Luca et PALLOTTA, Giuliana, 2015. Mining maritime vessel traffic: Promises, challenges, techniques. In : *OCEANS 2015 - Genova*. S.l. : s.n. mai 2015. p. 1-6.

CLARAMUNT, C., RAY, C., SALMON, L., CAMOSSI, E., HADZAGIC, M., JOUSSELME, A.-L., ANDRIENKO, G., ANDRIENKO, N., THEODORIDIS, Y. et VOUROS, G., 2017. Maritime data integration and analysis: Recent progress and research challenges. In : *Advances in Database Technology - EDBT*. 1 janvier 2017. Vol. 2017, p. 192-197. DOI 10.5441/002/edbt.2017.18.

Data Management Association (DAMA) UK Working Group on "Data Quality Dimensions", 2013. "The six primary dimensions for data quality assessment - Defining data quality dimensions (Final Version). Disponible à l'adresse : https://www.whitepapers.em360tech.com/wpcontent/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf

DENG, Feng, GUO, Sitong, DENG, Yong, CHU, Hanyue, ZHU, Qingmeng et SUN, Fuchun, 2014. Vessel track information mining using AIS data. In : *2014 International Conference on Multisensor Fusion and Information Integration for Intelligent Systems (MFI)*. S.l. : s.n. septembre 2014. p. 1-6.

DONATI, Jessica et FINEREN, Daniel, 2012. Reuters: Iran Falsifying AIS Data to Conceal Ship Movements. In : *gCaptain* [en ligne]. 6 décembre 2012. [Consulté le 26 juin 2020]. Disponible à l'adresse : https://gcaptain.com/iran-falsifying-ais-data-to-conceal-ship-movements/.

DRUEY, Guy, 2020. *Etude, conception, collecte, curation et évaluation d'un scraping de sites web liés au transport maritime pour améliorer la prédiction du fret de matières premières*. Genève : Haute école de gestion de Genève. Travail de Master

DUFOUR, Christine et LARIVIÈRE, Vincent, 2017. Page d'accueil du cours SCI6060 - Méthodes de recherche en sciences de l'information donné à l'École de bibliothéconomie et des sciences de l'information, Université de Montréal, dans le cadre du programme de maîtrise en sciences de l'information (2e cycle). cours.ebsi.umontreal.ca [en ligne]. 14 janvier 2016. Mis à jour le 19 décembre 2017 [Consulté le 05 juillet 2020]. Disponible à l'adresse : http://cours.ebsi.umontreal.ca/sci6060/

FABBRI, Tommaso, VICEN-BUENO, Raul, GRASSO, Raffaele, PALLOTTA, Giuliana, MILLEFIORI, Leonardo M. et CAZZANTI, Luca, 2015. Optimization of surveillance vessel network planning in maritime command and control systems by fusing METOC AIS vessel traffic information. In : *OCEANS 2015 - Genova*. S.l. : s.n. mai 2015. p. 1-7.

FELSKI, A. et JASKÓLSKI, K., 2012. Information unfitness of AIS. In : *Annual of Navigation* [en ligne]. 2012. Vol. No. 19, part 1. [Consulté le 25 avril 2020]. DOI 10.2478/v10367-012-0002-z. Disponible à l'adresse : http://yadda.icm.edu.pl/baztech/element/bwmeta1.element.baztech-ceb3fe6c-cd33-4cc7-aad8-e2967b17477e.

FILIPIAK, Dominik, STRÓŻYNA, Milena, WĘCEL, Krzysztof et ABRAMOWICZ, Witold, 2018. Big Data for Anomaly Detection in Maritime Surveillance: Spatial AIS Data Analysis for Tankers. In : *Zeszyty Naukowe Akademii Marynarki Wojennej*. 1 décembre 2018. Vol. 215, p. 5-28. DOI 10.2478/sjpna-2018-0024.

FORTIN, Marie-Fabienne et GAGNON, Johanne, 2016. Fondements et étapes du processus de recherche : Méthodes quantitatives et qualitatives. 3e éd. Montréal : Chenelière Éducation.

FOX, Christopher, LEVITIN, Anany et REDMAN, Thomas, 1994. The notion of data and its quality dimensions. In : *Information Processing & Management*. 28 février 1994. Vol. 30, p. 9-19. DOI 10.1016/0306-4573(94)90020-5.

FU, Qian et EASTON, John M., 2017. Understanding data quality: Ensuring data quality by design in the rail industry. In : *2017 IEEE International Conference on Big Data (Big Data)*. Décembre 2017. p. 3792-3799.

FÜRBER, Christian et HEPP, Martin, 2011. SWIQA – A SEMANTIC WEB INFORMATION QUALITY ASSESSMENT FRAMEWORK. In : *ECIS 2011 Proceedings* [en ligne]. 6 octobre 2011. n°76. Disponible à l'adresse : https://aisel.aisnet.org/ecis2011/76.

GERRITSEN, H. D., MINTO, C. et LORDAN, C., 2013. How much of the seabed is impacted by mobile fishing gear? Absolute estimates from Vessel Monitoring Systems (VMS) point data. In : [en ligne]. 2013. [Consulté le 6 juin 2020]. DOI DOI: 10.1093/icesjms/fst017. Disponible à l'adresse : https://oar.marine.ie/handle/10793/950.

GITZEL, Ralf, TURRIN, Simone, MACZEY, Sylvia, WU, Shaomin et SCHMITZ, Björn, 2016. A Data Quality Metrics Hierarchy for Reliability Data. In : *Proceedings of the 9thIMA International Conference on Modellingin Industrial Maintenance and Reliability* [en ligne]. London : Institute of Mathematics and its Applications. [Consulté le 25 avril 2020]. ISBN 978-0-905091-31-0. Disponible à l'adresse : https://kar.kent.ac.uk/56313/.

GITZEL, Ralf, TURRING, Simone et MACZEY, Sylvia, 2015. A Data Quality Dashboard for Reliability Data. In : *2015 IEEE 17th Conference on Business Informatics*. Juillet 2015. p. 90-97.

HARATI MOKHTARI, Abbas, WALL, Alan, BROOKS, Philip et WANG, Jin, 2007. Automatic Identification System (AIS): Data Reliability and Human Error Implications. In : *Journal of Navigation*. 1 septembre 2007. Vol. 60, p. 373-389. DOI 10.1017/S0373463307004298.

HARTIG, Olaf et ZHAO, Jun, 2009. Using web data provenance for quality assessment. In : *Proceedings of the First International Conference on Semantic Web in Provenance Management - Volume 526* [en ligne]. Washington DC : CEUR-WS.org. 25 octobre 2009. p. 29–34. [Consulté le 25 avril 2020]. Disponible à l'adresse : https://dl.acm.org/doi/10.5555/2889875.2889881.

IMO, 2015. Resolution A.1106(29): revised guidelines for the onboard operational use of shipborne automatic identification systems (AIS). Disponible à l'adresse : http://www.imo.org/en/KnowledgeCentre/IndexofIMOResolutions/Assembly/Documents/A.1106(29).pdf

IMO, 2004. International Convention for the Safety of Life at Sea. International Maritime Organization. www.imo.org/

IMO, 2008. International Convention for the Safety of Life at Sea. International Maritime Organization. www.imo.org/

IMO, 2015. Resolution A.1106(29): revised guidelines for the onboard operational use of shipborne automatic identification systems (AIS). Disponible à l'adresse : http://www.imo.org/en/KnowledgeCentre/IndexofIMOResolutions/Assembly/Documents/A.1106(29).pdf

IPHAR, C., NAPOLI, A. et RAY, C., 2015. Data Quality Assessment for Maritime Situation Awareness. In : *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*. 1 août 2015. Vol. 3, p. 291-296. DOI 10.5194/isprsannals-II-3-W5-291-2015.

IPHAR, Clément, NAPOLI, Aldo et RAY, Cyril, 2015. Detection of false AIS messages for the improvement of maritime situational awareness. In : *OCEANS 2015 - MTS/IEEE Washington*. S.l. : s.n. octobre 2015. p. 1-7.

IPHAR, Clément, NAPOLI, Aldo et RAY, Cyril, 2016a. A method for integrity assessment of information in a worldwide maritime localization system. In : *19th AGILE International Conference on Geographic Information Science (AGILE 2016)* [en ligne]. S.l. : s.n. 14 juin 2016. [Consulté le 25 avril 2020]. Disponible à l'adresse : https://hal-mines-paristech.archives-ouvertes.fr/hal-01421920.

IPHAR, Clément, NAPOLI, Aldo et RAY, Cyril, 2016b. On the Interest of Data Mining for an Integrity Assessment of AIS Messages. In : *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. S.l. : s.n. décembre 2016. p. 368-373.

IPHAR, Clément, NAPOLI, Aldo et RAY, Cyril, 2017. Integrity Assessment of a Worldwide Maritime Tracking System for a Geospatial Analysis at Sea. In : *20th AGILE International Conference on Geographic Information Science (AGILE 2017)* [en ligne]. Wageningen, Netherlands : Societal Geo-innovation. mai 2017. p. 4 pages. [Consulté le 25 avril 2020]. Disponible à l'adresse : https://hal-mines-paristech.archives-ouvertes.fr/hal-01534116.

IPHAR, Clément, RAY, Cyril et NAPOLI, Aldo, 2020. Data integrity assessment for maritime anomaly detection. In : *Expert Systems with Applications*. 1 juin 2020. Vol. 147, p. 113219. DOI 10.1016/j.eswa.2020.113219.

JESIĻEVSKA, Svetlana, 2017. Data Quality Dimensions to Ensure Optimal Data Quality. In : *The Romanian Economic Journal* [en ligne]. 2017. [Consulté le 30 juillet 2020]. Disponible à l'adresse : /paper/Data-Quality-Dimensions-to-Ensure-Optimal-Data-Jesi%C4%BCevska/ad0428b47df9f045392df511977dfe77f8a03dc5.

KALUZA, Pablo, KOELZSCH, Andrea, GASTNER, Michael et BLASIUS, Bernd, 2010. The complex network of global cargo ship movement. In : *Journal of the Royal Society, Interface / the Royal Society*. 6 juillet 2010. Vol. 7, p. 1093-103. DOI 10.1098/rsif.2009.0495.

KALYVAS, Christos, KOKKOS, Athanasios et TZOURAMANIS, Theodoros, 2017. A survey of official online sources of high-quality free-of-charge geospatial data for maritime geographic information systems applications. In : *Information Systems*. 1 avril 2017. Vol. 65, p. 36-51. DOI 10.1016/j.is.2016.11.002.

KAZEMI, Samira, ABGHARI, Shahrooz, LAVESSON, Niklas, JOHNSON, Henric et RYMAN, Peter, 2013. Open Data for Anomaly Detection in Maritime Surveillance. In : *Expert Systems with Applications*. 4 mai 2013. Vol. 40. DOI 10.1016/j.eswa.2013.04.029.

KOS, Serdjo, VUKIĆ, Mate et BRČIĆ, David, 2013. Use of universal protocol for entering the port of destination in AIS device. In : *International Maritime Science Conference* [en ligne]. Solin : s.n. 23 avril 2013. Disponible à l'adresse : https://www.researchgate.net/publication/236723236_Use_of_universal_protocol_for_entering_the_port_of_destination_in_AIS_device.

LAST, Philipp, BAHLKE, Christian, HERING-BERTRAM, Martin et LINSEN, Lars, 2014. Comprehensive Analysis of Automatic Identification System (AIS) Data in Regard to Vessel Movement Prediction. In : *The Journal of Navigation*. septembre 2014. Vol. 67, n° 5, p. 791-809. DOI 10.1017/S0373463314000253.

LEVY, Yair et ELLIS, Timothy J., [sans date]. A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. In : *Informing Science: The International Journal of an Emerging Transdiscipline*. Vol. 9, p. 181-212.

LOSHIN, David, 2006. *Informatica_Whitepaper_Monitoring_DQ_Using_Metrics.pdf* [en ligne]. [Consulté le 15 mai 2020 a]. Disponible à l'adresse : https://it.ojp.gov/documents/Informatica_Whitepaper_Monitoring_DQ_Using_Metrics.pdf.

MAŁYSZKO, Jacek, ABRAMOWICZ, Witold et STRÓŻYNA, Milena, 2016. Named Entity Disambiguation for Maritime-related Data Retrieved from Heterogenous Sources. In : *TransNav - The International Journal on Marine Navigation and Safety of Sea Transportation*. 1 septembre 2016. Vol. 10, p. 465-477. DOI 10.12716/1001.10.03.12.

MAO, Shangbo, TU, Enmei, ZHANG, Guanghao, RACHMAWATI, Lily, RAJABALLY, Eshan et HUANG, Guang-Bin, 2018. An Automatic Identification System (AIS) Database for Maritime Trajectory Prediction and Data Mining. In : CAO, Jiuwen, CAMBRIA, Erik, LENDASSE, Amaury, MICHE, Yoan et VONG, Chi Man (éd.), *Proceedings of ELM-2016*. Cham : Springer International Publishing. 2018. p. 241-257.

MAREV, Milen, COMPATANGELO, Ernesto et VASCONCELOS, Wamberto, 2018. *Towards a context-dependent numerical data quality evaluation framework*. Disponible à l'adresse : https://arxiv.org/abs/1810.09399v1

MAZZARELLA, Fabio, VESPE, Michele, ALESSANDRINI, Alfredo, TARCHI, Dario, AULICINO, Giuseppe et VOLLERO, Antonio, 2017. A novel anomaly detection approach to identify intentional AIS on-off switching. In : *Expert Systems with Applications*. 15 juillet 2017. Vol. 78, p. 110-123. DOI 10.1016/j.eswa.2017.02.011.

McGilvray, D., Ten Steps to Quality Data and Trusted Information, 1 er éd., Morgan Kaufmann, USA, 2008, 325 p.

MERINO, Jorge, CABALLERO, Ismael, RIVAS, Bibiano, SERRANO, Manuel et PIATTINI, Mario, 2016. A Data Quality in Use model for Big Data. In : *Future Generation Computer Systems*. 1 octobre 2016. Vol. 63, p. 123-130. DOI 10.1016/j.future.2015.11.024.

PAN, Zheng et DENG, Shujun, 2009. Vessel Real-Time Monitoring System Based on AIS Temporal Database. In : *2009 International Conference on Information Management, Innovation Management and Industrial Engineering*. Décembre 2009. p. 611-614.

PARLEMENT EUROPÉEN ET DU CONSEIL, mars 2009. *PDF.pdf* [en ligne]. S.l. : s.n. [Consulté le 18 mai 2020 c]. Disponible à l'adresse : https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32009R0223&from=en.

PARSE HUB, 2020. Web Scraping VS Data Mining : What's the Difference? [enregistrement vidéo]. *YouTube* [en ligne]. 16 mars 2020. Consulté le 4 juillet 2020]. Disponible à l'adresse : https://www.youtube.com/watch?v=ssUeAe30UTI

PAWAR, Supriya Haribhau, 2017. An Assessment Model to Evaluate Quality Attributes In Big Data Quality. In : . 2017. Vol. 5, n 2, p. 4.

PIPINO, Leo L., LEE, Yang W. et WANG, Richard Y., 2002. Data quality assessment. In : *Communications of the ACM*. 1 avril 2002. Vol. 45, n 4, p. 211–218. DOI 10.1145/505248.506010.

RĂDULESCU, Dan, ST-HILAIRE, Marie-Odette, ALLARD, Yannick et HAMMOND, Tim, 2016. Sharing AIS Related Anomalies (SARA). In : [en ligne]. 2016. [Consulté le 30 avril 2020]. Disponible à l'adresse : https://www.semanticscholar.org/paper/Sharing-AIS-Related-Anomalies-(SARA)-R%C4%83dulescu-St-Hilaire/dbcb37c1b9874cead2bcb0597d7c21401d93d9d7.

RAY, Cyril, 2018. Data Variety and Integrity Assessment for Maritime Anomaly Detection. In : *BDCSIntell*. 2018.

RAY, Cyril, GALLEN, Romain, IPHAR, Clément, NAPOLI, Aldo et BOUJU, Alain, 2015. DeAIS project: Detection of AIS spoofing and resulting risks. In : *OCEANS 2015 - Genova*. Mai 2015. p. 1-6.

RAY, Cyril, IPHAR, Clément et NAPOLI, Aldo, 2016. Methodology for Real-Time Detection of AIS Falsification. In : *Maritime Knowledge Discovery and Anomaly*

Detection Workshop [en ligne]. S.l. : s.n. 5 juillet 2016. p. 74-77-ISBN 978-92-79-61301-2. [Consulté le 25 avril 2020]. Disponible à l'adresse : https://hal-mines-paristech.archives-ouvertes.fr/hal-01421910.

REDOUTEY, Martin, SCOTTI, Eric, JENSEN, Christian, RAY, Cyril et CLARAMUNT, Christophe, 2008. Efficient Vessel Tracking with Accuracy Guarantees. In : BERTOLOTTO, Michela, RAY, Cyril et LI, Xiang (éd.), *Web and Wireless Geographical Information Systems*. Berlin, Heidelberg : Springer. 2008. p. 140-151.

SCANNAPIECO, Monica, MISSIER, Paolo et BATINI, Carlo, 2005. Data Quality at a Glance. In : *Datenbank-Spektrum*. 1 janvier 2005. Vol. 14, p. 6-14.

SERRY, Arnaud et LÉVÊQUE, Laurent, 2015. Le système d'identification automatique (AIS). Une source de données pour étudier la circulation maritime. In : *Netcom. Réseaux, communication et territoires*. 14 décembre 2015. n 29-1/2, p. 177-202. DOI 10.4000/netcom.1943.

SHELMERDINE, Richard L., 2015. Teasing out the detail: How our understanding of marine AIS data can better inform industries, developments, and planning. In : *Marine Policy*. 1 avril 2015. Vol. 54, p. 17-25. DOI 10.1016/j.marpol.2014.12.010.

STEIDEL, Matthias, LAMM, Arne, FEUERSTACK, Sebastian et HAHN, Axel, 2019. Correcting the Destination Information in Automatic Identification System Messages. In : ABRAMOWICZ, Witold et CORCHUELO, Rafael (éd.), *Business Information Systems Workshops*. Cham : Springer International Publishing. 2019. p. 496-507.

STRONG, Diane, LEE, Yang et WANG, Richard, 2002. Data Quality in Context. In : *Communications of the ACM*. 11 août 2002. Vol. 40. DOI 10.1145/253769.253804.

STRÓŻYNA, Milena, EIDEN, Gerd, ABRAMOWICZ, Witold, FILIPIAK, Dominik, MAŁYSZKO, Jacek et WĘCEL, Krzysztof, 2018. A framework for the quality-based selection and retrieval of open data - a use case from the maritime domain. In : *Electronic Markets*. 1 mai 2018. Vol. 28, n 2, p. 219-233. DOI 10.1007/s12525-017-0277-y.

STRÓŻYNA, Milena, MAŁYSZKO, Jacek, WĘCEL, Krzysztof, FILIPIAK, Dominik et ABRAMOWICZ, Witold, 2016. Architecture of Maritime Awareness System Supplied with External Information. In : *Annual of Navigation*. 31 décembre 2016. Vol. 23, p. 135-149. DOI 10.1515/aon-2016-0009.

TU, Enmei, ZHANG, Guanghao, RACHMAWATI, Lily, RAJABALLY, Eshan et HUANG, Guang-Bin, 2018. Exploiting AIS Data for Intelligent Maritime Navigation: A Comprehensive Survey From Data to Methodology. In : *IEEE Transactions on Intelligent Transportation Systems*. mai 2018. Vol. 19, n 5, p. 1559-1582. DOI 10.1109/TITS.2017.2724551.

Tunaley, J.K.E., 2013. Utility of Various AIS Messages for Maritime Awareness. Presented at the 8th ASAR Workshop, Longueuil, Canada.

UNION INTERNATIONALE DES TELECOMMUNICATIONS, Février 2014. Caractéristiques techniques d'un système d'identification automatique utilisant l'accès multiple par répartition dans le temps et fonctionnant dans la bande de fréquences attribuée aux services mobiles maritimes en ondes métriques. *R-REC-M.1371-5-201402-I!!PDF-F.pdf* [en ligne]. [Consulté le 1 juin 2020 d]. Disponible à l'adresse : https://extranet.itu.int/brdocsearch/R-REC/R-REC-M/R-REC-M.1371/R-REC-M.1371-5-201402-I/R-REC-M.1371-5-201402-I!!PDF-F.pdf.

VAZIRI, Reza, MOHSENZADEH, Mehran et HABIBI, Jafar, 2019. Measuring data quality with weighted metrics. In : *Total Quality Management & Business Excellence*. 3 avril 2019. Vol. 30, n 5-6, p. 708-720. DOI 10.1080/14783363.2017.1332954.

VETRÒ, Antonio, CANOVA, Lorenzo, TORCHIANO, Marco, MINOTAS, Camilo Orozco, IEMMA, Raimondo et MORANDO, Federico, 2016. Open data quality measurement

framework: Definition and application to Open Government Data. In : *Government Information Quarterly*. 1 avril 2016. Vol. 33, n 2, p. 325-337. DOI 10.1016/j.giq.2016.02.001.

WAND, Yair et WANG, Richard Y., 1996. Anchoring data quality dimensions in ontological foundations. In : *Communications of the ACM*. 1 novembre 1996. Vol. 39, n 11, p. 86–95. DOI 10.1145/240455.240479.

WANG, Richard Y. et STRONG, Diane M., 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. In : *Journal of Management Information Systems*. 1 mars 1996. Vol. 12, n 4, p. 5-33. DOI 10.1080/07421222.1996.11518099.

WEBMASTER, [sans date]. EMSA Facts & Figures 2019. In : [en ligne]. [Consulté le 26 juin 2020]. Disponible à l'adresse : http://www.emsa.europa.eu/emsa-documents/latest/item/3900-emsa-facts-figures-2019.html.

Windward, 2014. AIS Data on the Hish Seas: An Analysis of the Magnitude and Implications of Growing Data Manipulation at Sea, Windward, October 2014. Disponible à l'adresse : http://maritime-connector.com/news/general/ais-data-on-the-high-seas-an-analysis-of-the-magnitude-and-implications-of-growing-data-manipulation-at-sea/

WINTHER, Morten, CHRISTENSEN, Jesper, PLEJDRUP, Marlene, RAVN, Erik, ERIKSSON, Omar et KRISTENSEN, Hans, 2014. Emission inventories for ships in the arctic based on satellite sampled AIS data. In : *Atmospheric Environment*. 1 juillet 2014. Vol. 91, p. 1–14. DOI 10.1016/j.atmosenv.2014.03.006.

YANG, Dong, WU, Lingxiao, WANG, Shuaian, JIA, Haiying et LI, Kevin X., 2019. How big data enriches maritime research – a critical review of Automatic Identification System (AIS) data applications. In : *Transport Reviews*. 2 novembre 2019. Vol. 39, n 6, p. 755-773. DOI 10.1080/01441647.2019.1649315.

Annexe 1: Summary of AIS Messages

Tableau 16 : Summary of AIS Messages

Category	Message	Description
Standard	1	Scheduled position report (class A)
	2	Assigned position report (class A)
	3	Special position report (class A)
	5	Static report (class A)
	9	SAR aircraft position report
	18	Position report (class B)
	19	Extended position report (class B)
	24	Static report (classB)
	27	Long range position report
AToN	21	AToN report
Timing	4	Base station report
	10	UTC inquiry
	11	UTC response
Safety	12	Addressed text message
	13	Acknowledgment
	14	Broadcast text message
Binary	6	Addressed binary
	7	Binary acknowledgment

	8	Broadcast binary
	17	GNSS update
	25	Short binary (no acknowledgement)
	26	Binary with communications state
Other	15	Interrogation for specific messages
	16	Assignment mode command
	20	Data link management
	22	Channel management
	23	Group assignment command

(Tunaley 2013, p. 2)

Annexe 2 : Synthèse des attributs des données AIS selon le type d'information et la provenance de l'information

Tableau 17 : Synthèse des données AIS

Data Field	Type	Description	Information generation
AIS identity and location	Static	Maritime Mobile Service Identity (MMSI) and the location of the system's antenna on board	Set on installation (Note that this might need amending if the ship changes ownership)
Ship identity		Ship name, IMO number, type, and call sign of the ship	Set on installation (Note that this might need amending if the ship changes ownership) Type: Select from pre-installed list
Ship size		Length and width of the ship	Set on installation or if changed
Ship position	Dynamic	Latitude and longitude (up to 0.0001 min accuracy)	Automatically updated from the position sensor connected to AIS. The accuracy indication is approximately 10 m.
Speed		Ranging from 0 knot to 102 knots (0.1 knot resolution)	Automatically updated from the position sensor connected to AIS. This information might not be available
Rate of turn		Right or left (ranging from 0 to 720° per minute)	Automatically updated from the ship's ROT sensor or derived from the gyro. This information might not be available
Navigation direction		Shipping course, heading, and bearing of the ship	Automatically updated from the ship's heading sensor connected to AIS
Time stamp		Second field of the UTC time when the subject data packet was generated	Automatically updated from ship's main position sensor connected to AIS
Navigation status		Includes "at anchor,""under way using engine(s)," and "not under command"	Navigational status information has to be manually entered by the

Destination and ETA	Voyage related	Destination port and the estimated time of arrival of the ship	To be manually entered at the start of the voyage and kept up to date as necessary
Route plan (waypoints)			To be manually entered at the start of the voyage, at the discretion of the master, and updated when required
Draught		Ranges from 0.1 m to 25.5 m	To be manually entered at the start of the voyage using the maximum draft for the voyage and amended as required (p. ex. – result of de-ballasting prior to port entry)
Hazardous cargo (type)		• DG (Dangerous goods) • HS (Harmful substances) • MP (Marine pollutants)	To be manually entered at the start of the voyage confirming whether or not hazardous cargo is being carried. Indications of quantities are not required
Short safety-related messages	Safety-related	Free format	Free format short text messages would be manually entered, addressed either a specific addressee or broadcast to all ships and shore stations

(Yang & al. 2019, p. 758 et IMO Resolution A.1106 (29) 2015, p. 5-6)

Annexe 3 : Valeurs acceptables pour les informations statiques

Le numéro MMSI des navires a été spécifié par l'ITU. Ce dernier est composé de 9 chiffres compris entre 201000000 et 775999999 (MIDXXXXXX). Les trois premiers chiffres de ce dernier sont le MID et représentent un code relatif à un pays

Tableau 18 : Examples of MID country codes[28]

Code (MID)	Country
710	Brésil
316	Canada
725	Chili
219, 220	Danemark
237, 239-241	Grèce
251	Islande
440, 441	Corée du Sud
636, 637	Libéria
257-259	Norvège
351-357, 370-373	Panama
338, 366-369	USA

(Tunaley 2013, p. 3)

[28] Tous les MID sont librement disponibles et consultables à l'adresse suivante : https://www.itu.int/en/ITU-R/terrestrial/fmd/Pages/mid.aspx

Annexe 4 : Valeurs acceptables pour les informations dynamiques

Tableau 19 : Intervalles pour les équipements mobiles de navire de classe A

Description	Intervalles (secondes)
Navire à l'ancre ou au mouillage et ne se déplaçant pas à plus de 3 noeuds	180 s (3 min)
Navire à l'ancre ou au mouillage et se déplaçant à plus de 3 noeuds	10 s
Navire à 0-14 nœud (SOG)	10 s
Navire à 0-14 noeud et changeant de route (SOG)	3 1/3 s
Navire à 14-23 nœuds (SOG)	6 s
Navire à 14-23 noeuds et changeant de route (SOG)	2 s
Navire à plus de 23 nœuds (SOG)	2 s
Navire à plus de 23 noeuds et changeant de route (SOG)	2 s

(ITU-R M.1371-5 2014, p. 9)

Tableau 20 : Valeurs numériques valides en lien avec les informations dynamiques[29]

Data Field	Unit	Range	n.a. value
Longitude	[°]	±180	181
Latitude	[°]	±90	91
Rate of turn (ROT)	$[\frac{°}{m}]$	±127	128
Speed over ground (SOG)	[kn]	[0,1022]	1023
Course over ground (COG)	[°]	[0,3599]	3600
Heading (HDG)	[°]	[0,359]	511
Position accuracy (ACC)[30]	-	[true, false]	false

[29] Données issues des différents capteurs directement reliés au système AIS sur le navire

[30] Position accuracy is represented as a flag without a unit. It indicates whether the received position of a vessel has an accuracy of ACC ≤10m.

(Last & al. 2014, p.795)

Tableau 21 : Valeurs possibles pour le statut de navigation[31]

Code	Description (FR)	Description (EN)
0	En route au moteur	Under way using engine
1	À l'ancre	At anchor
2	Non manoeuvrable	Not under command
3	Manoeuvrabilité réduite	Restricted maneuverability
4	Limité par son tirant d'eau	Constrained by her draught
5	Au mouillage	Moored
6	Échoué	Aground
7	Pêche	Engaged in fishing
8	Navigation à la voile	Under way sailing
9	Réservé pour une modification future du statut de navigation pour des navires transportant des DG, HS ou MP à risques ou des polluants de la catégorie C de l'OMI (HSC)	Reserved for future amendment of navigational status for ships carrying DG, HS, or MP, or IMO hazard or pollutant category C, high-speed craft (HSC)
10	Réservé pour une modification future du statut de navigation pour des navires transportant des DG, HS ou MP à risques ou des polluants de la catégorie A de l'OMI (WIG)	Reserved for future amendment of navigational status for ships carrying dangerous goods (DG), harmful substances (HS) or marine pollutants (MP), or IMO hazard or pollutant category A, wing in ground (WIG)
11	Remorquage de navires à propulsion mécanique vers l'arrière (utilisation régionale)	Power-driven vessel towing astern (regional use)
12	Poussage ou remorquage à couple de navires à propulsion mécanique (utilisation régionale)	Power-driven vessel pushing ahead or towing alongside (regional use)

[31] Ce dernier est configuré manuellement par l'équipage

13	Réservé pour utilisation future	Reserved for future use
14	Recherche et sauvetage AIS (AIS-SART) (active); personne à la mer AIS (MOB-AIS) et radiobalise de localisation des sinistres AIS (RLS-AIS)	AIS-SART (active), MOB-AIS, EPIRB-AIS
15	Non défini = par défaut (aussi utilisé par le AIS-SART soumis à des essais, le MOB-AIS et RLS-AIS).	Undefined = default (also used by AIS-SART, MOB-AIS and EPIRB-AIS under test)

(ITU-R M.1371-5 2014, p. 119)

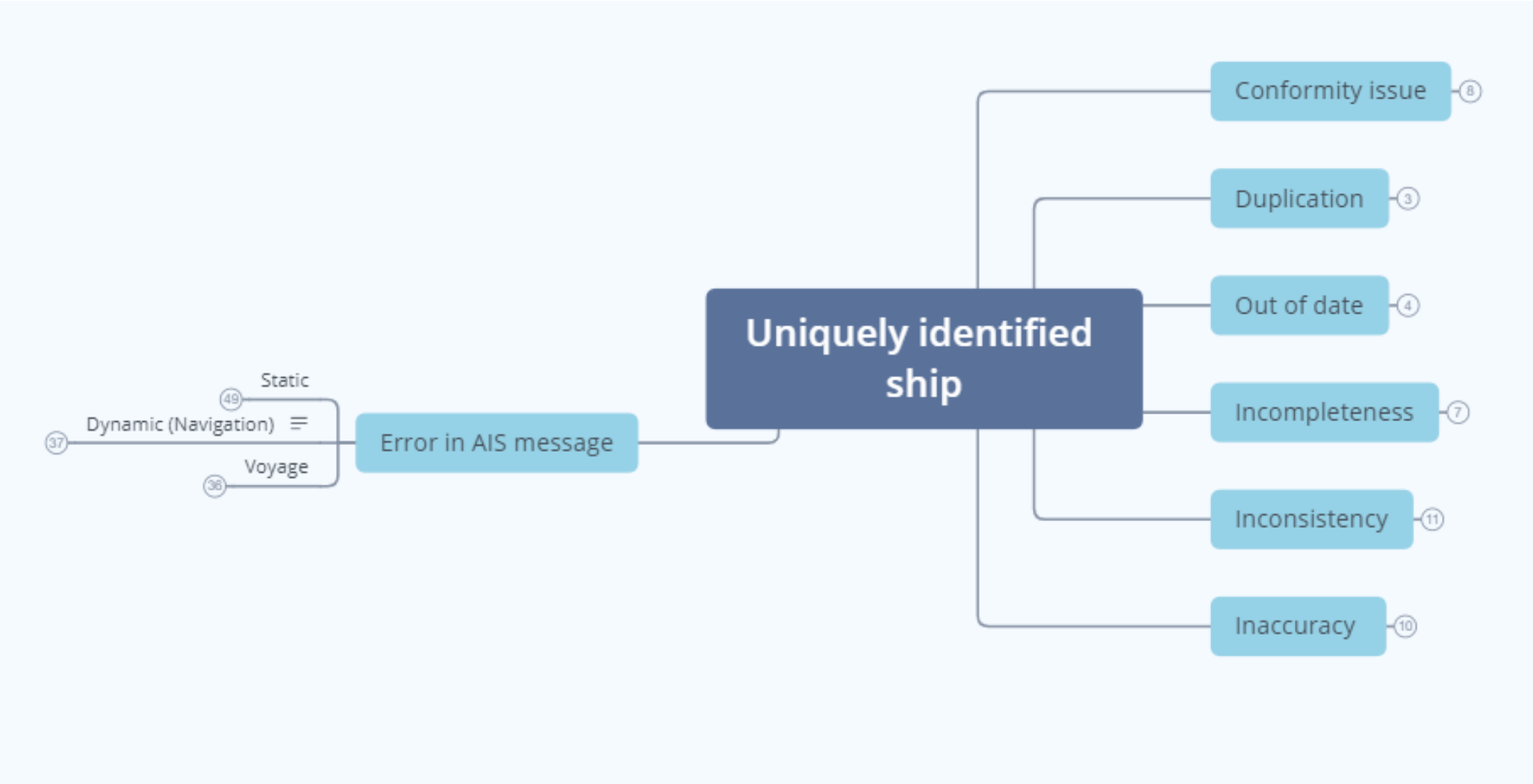
Annexe 5 : Valeurs acceptables pour les informations liées au voyage

Tableau 22 : Recommandation de l'IMO sur l'utilisation du UN/LOCODE

Norme No	Message	Description
1	DE HAM>NL RTM	Hamburg to Rotterdam.
2	DE HAM>?? ???	Hamburg to unknown destination
3	XX XXX>DE HAM	Unknown origin to Hamburg.
4	===Orrviken	If the destination does not have a UN/LOCODE, "===" should be entered, followed by the English name of the destination.
5	DE HAM> === US WC	If the destination is a general area, the known name or accepted abbreviation of the area should be used.

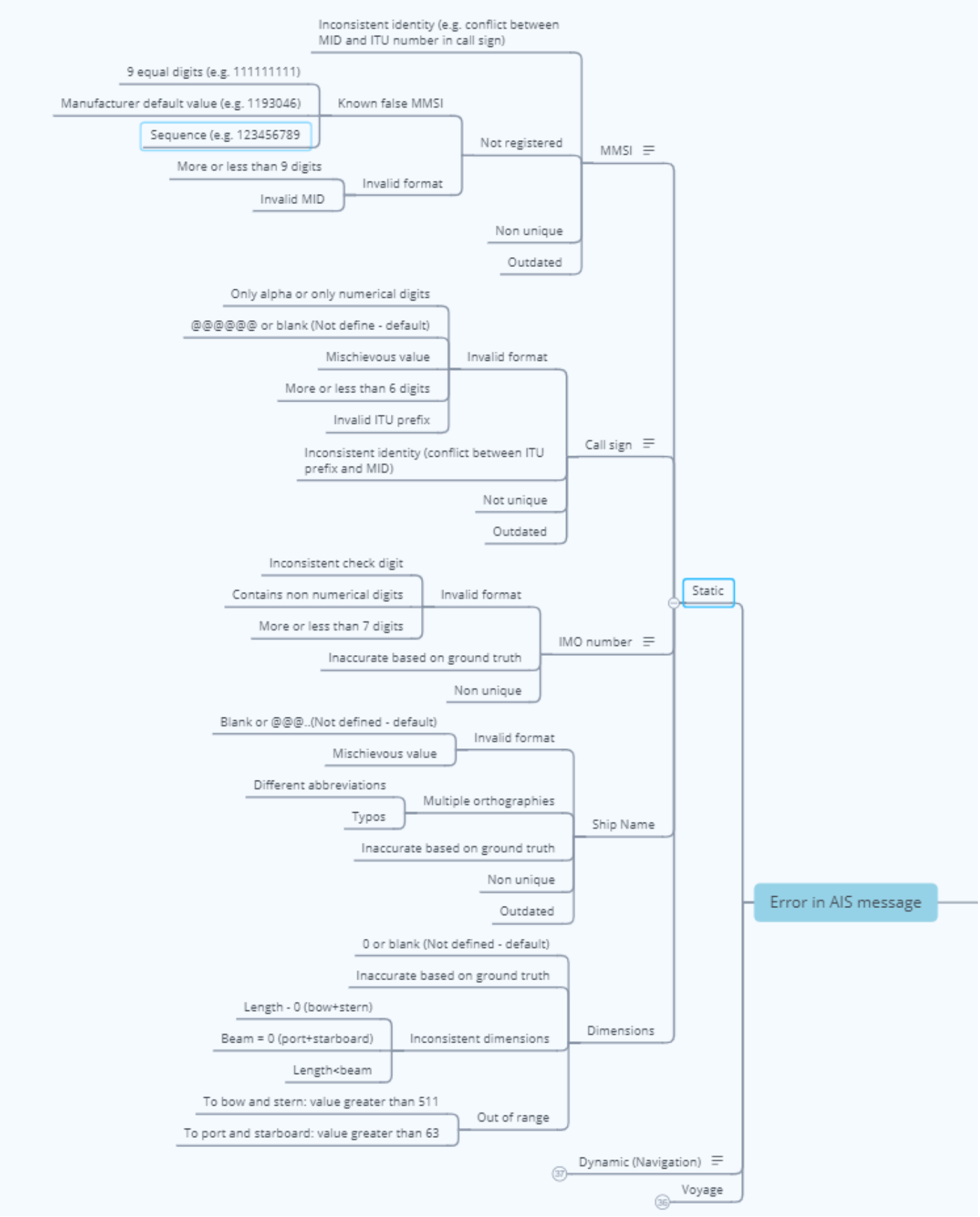
(Steidel & al. 2019, p. 5)

Annexe 6 : Taxonomie générale des anomalies AIS



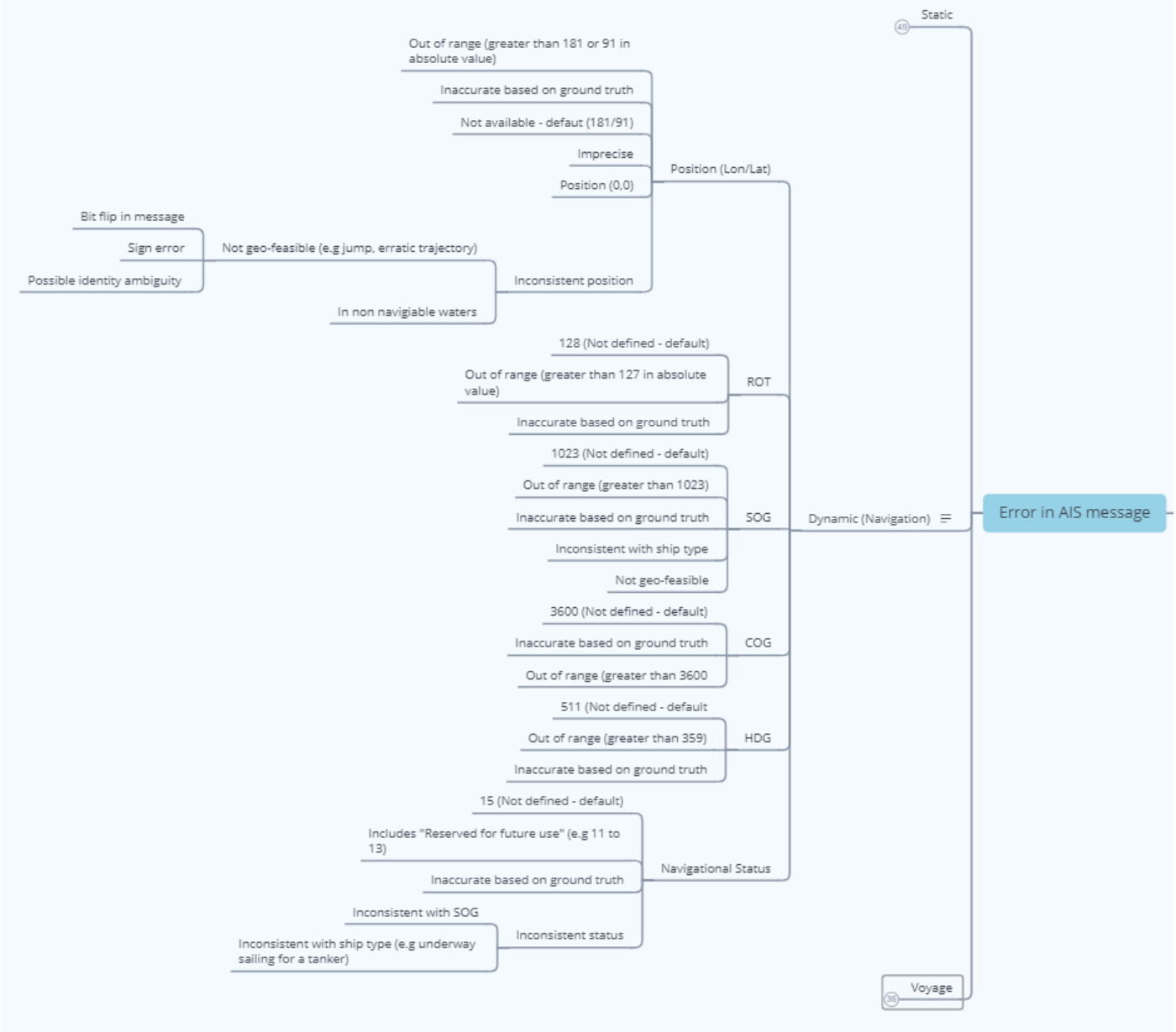
(RĂDULESCU & al. 2016, p. 16)

Annexe 7 : Anomalies AIS 1er niveau, erreurs champs « statique »



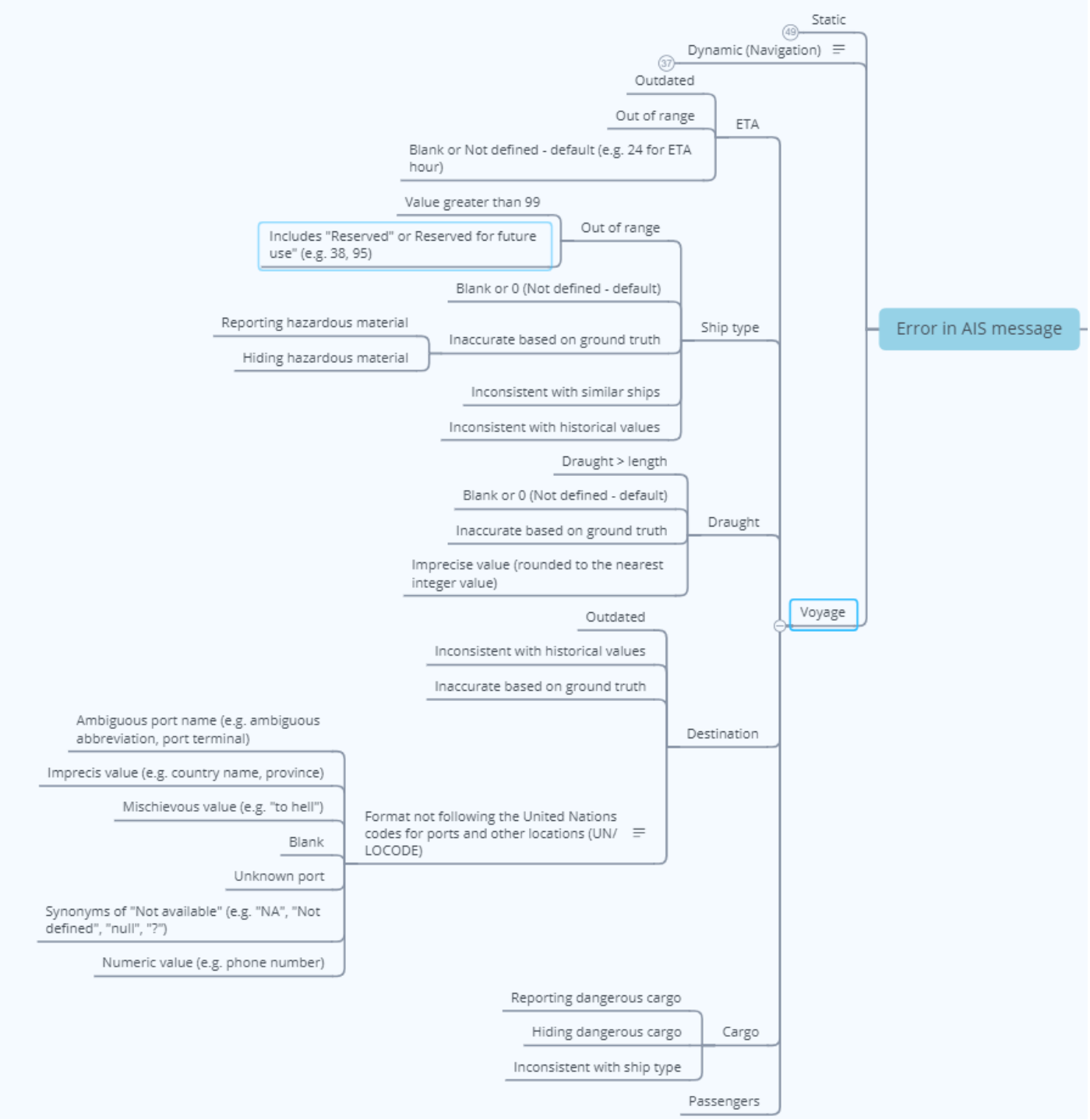
(RĂDULESCU & al. 2016, p. 18)

Annexe 8 : Anomalies AIS 1er niveau, erreurs champs « dynamique »



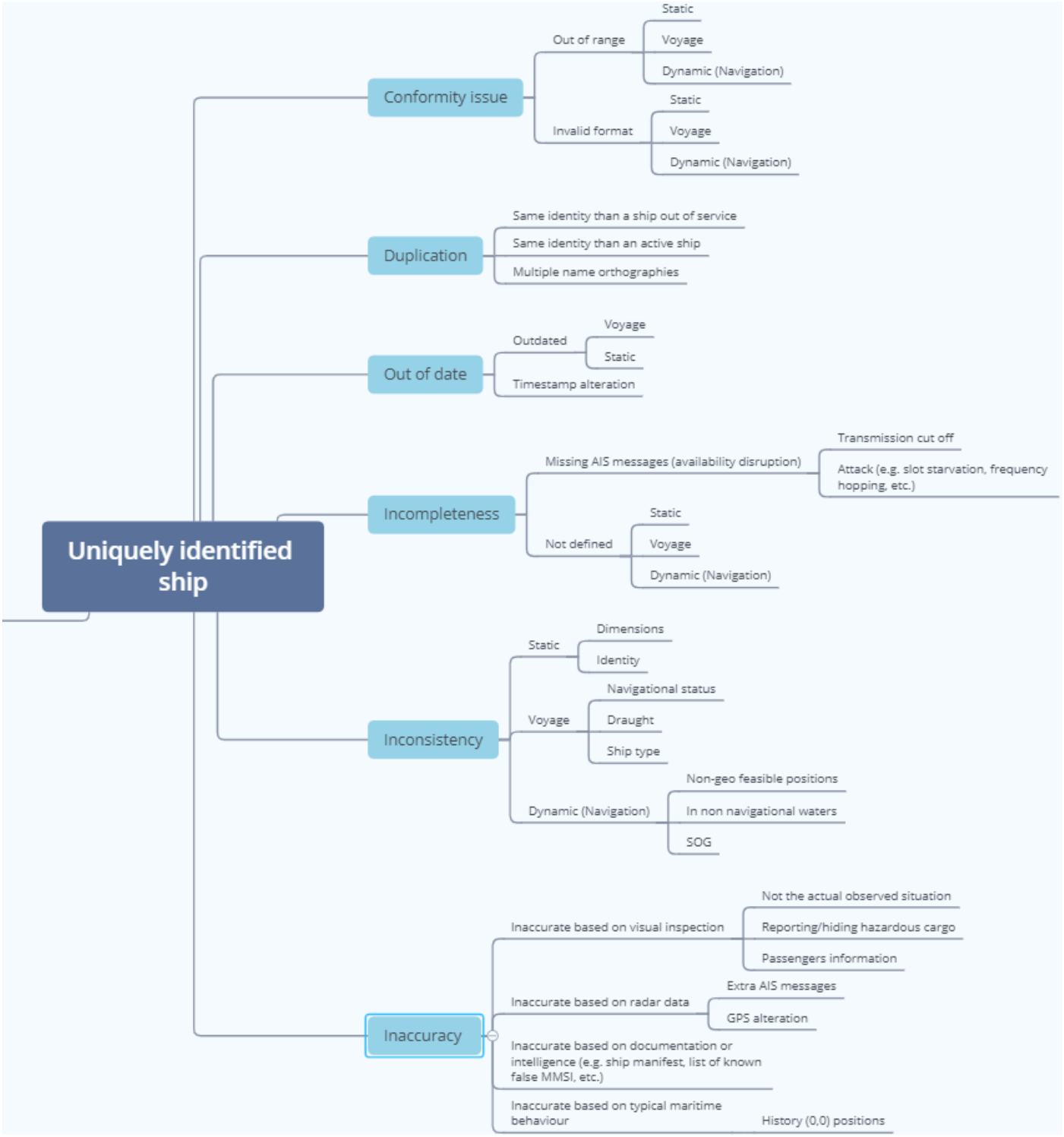
(RĂDULESCU & al. 2016, p. 20)

Annexe 9 : Anomalies AIS 1er niveau, erreurs champs « voyage »



d(RĂDULESCU & al. 2016, p. 19)

Annexe 10 : Taxonomie des anomalies AIS 2ème niveau



(RĂDULESCU & al. 2016, p. 21)

Annexe 11 : Dimensions de la qualité dans la littérature

Tableau 23 : Dimensions de la qualité dans la littérature et leur définition

Dimension agrégée	Dimension de l'auteur	Définition	Auteur	Fréquence
Completeness **Complétude/Exhaustivité**	Completeness	The proportion of stored data against the potential of "100% complete"	DAMA (2013)	18
	Reliability	Les auteurs ont choisi d'agréger avec cette dimension avec la dimension de précision (Accuracy)	Stróżyna et al. (2007)	
	Completeness	Completeness means that the values of all components of a dataset are valid.	Cai & Zhu (2015)	
	Completeness	Check the extracted data is not missing	Pawar (2017)	
	Completeness	The degree to which a data collection has values for all attributes of all entities that are supposed to have values.	Fox et al. (1994)	
	Complétude	Quantité de valeurs renseignées	Berti-Equille (2004)	
	Data coverage	A measure of the availability and comprehensiveness of data compared to the total data universe or population of interest	McGilvray (2008)	

	Completeness	Percentage of complete cells: Indicates the percentage of complete cells in a dataset. It means the cells that are not empty and have a meaningful value assigned (i.e. a value coherent with the domain of the column). Percentage of complete rows: Indicates the percentage of complete rows in a dataset. It means the rows that don't have any incomplete cell.	Vetro et al. (2016)	
	Completeness	Completeness may be assessed by the definition of mandatory property and literal value rules	Fürber & Hepp (2011)	
	Completeness	Not missing data	Gitzel et al, (2015)	
	Completeness	An expectation of completeness indicates that certain attributes should be assigned values in a data set.	Loshin (2006)	
	Completeness	The proportion of stored data against the potential of ‘100% complete’	Wei (2016)	
	Completeness	Any gaps in a numeric dataset may impact on the result of computational analytics based on averages, best fits, or regressions	Marev et al (2018)	
	Completeness	Completeness means that all of the required classes and properties should be represented while completeness at the	Behkamal et al. (2014)	

		data level refers to the missing values of properties with respect to the schema.		
	Completeness Appropriate amount of data	The extent to which data are of sufficient breadth, depth, and scope for the task at hand. The extent to which the quantity or volume of available data is appropriate	Wang et Strong (1996)	
	Completeness	Completeness can be generically defined as the extent to which data are of sufficient breadth, depth and scope for the task at hand.	Scannapieco et al. (2005)	
	Completeness	Definition of all the fields that should be filled in depending on the message	Iphar et al. (2015)	
	Completeness	Complete or not complete	Batini et al. (2009)	
	Completeness Appropriate amount of data	The extent to which data is not missing and is of sufficient breadth and depth for the task at hand. The extent to which the quantity or volume of available data is appropriate for the task at hand	Pipino et al. (2002)	
Uniqueness **Unicité/Duplication**	Uniqueness	Nothing will be recorded more than once based upon how that thing is identified. It is the inverse of an assessment of the level of duplication	DAMA (2013)	9

	Duplication	Duplication occurs when a real world entity is stored twice or more in a data source.	Scannapieco et al. (2005)	
	Duplication	A measure of unwanted duplication existing within or across systems for a particular field, record, or data set	McGilvray (2008)	
	Uniqueness	We understand uniqueness as the degree to which data is free of redundancies in breadth, depth, and scope	Fürber & Hepp (2011)	
	Uniqueness	Check that an element in the dataset is actually unique and that there are no repeating parameters	Marev et al (2018)	
	Uniqueness	Nothing will be recorded more than once based upon how it is identified.	Wei (2016)	
	Uniqueness	Defined as the non-redundancy characteristic of the entities, classes, properties, and values of properties in a dataset.	Behkamal et al. (2014)	
	Uniqueness	Asserting uniqueness of the entities within a data set implies that no entity exists more than once within the data set	Loshin (2006)	
	Duplicate records	A data collection may also contain extra triples, possibly with distinct values for the same attribute of the same entity	Fox et al. (1994)	
Timeliness	Timeliness	The degree to which data represent reality from the required point in time	DAMA (2013)	18

Actualisation/disponibilité	Timeliness Punctuality	Timeliness: the period between the availability of the information and the event or phenomenon it describes Punctuality: the delay between the date of the release of the data and the target date	Stróżyna et al. (2007)	
	Timeliness	Timeliness is defined as the time delay from data generation and acquisition to utilization.	Cai & Zhu (2015)	
	Timeliness and Availability	A measure of the degree to which data are current and available for use as specified and in the time frame in which they are expected.	McGilvray (2008)	
	Timeliness	The extracted input data is not old data. The timestamp is necessary when retrieving the data.	Pawar (2017)	
	Timeliness	The extent to which the age of the data is appropriate for the task at hand.	Wang et Strong (1996)	
	Currentness	Percentage of current rows: Indicates the percentage of rows of a dataset that have current values, it means that they don't have any value that refers to a previous or a following period of time. Delay in publication: Indicates the ratio between the delay in the publication (number of days passed between the moment in which the information is available and the publication of the	Vetro et al. (2016)	

	Expiration	dataset) and the period of time referred by the dataset (week, month, year). Delay after expiration: Indicates the ratio between the delay in the publication of a dataset after the expiration of its previous version and the period of time referred by the dataset (week, month, year).		
	Timeliness Currency	Timeliness can be measured as the time between when information is expected and when it is readily available for use Currency refers to the degree to which information is current with the world that it models	Loshin (2006)	
	Timeliness	The extent to which the data is sufficiently up-to-date for the task at hand	Pipino et al. (2002)	
	Currentness	A data is said to be *current* or *up-to-date* at time *t* if it is correct at time *t.* Remarque: les auteurs définissent encore la notion « age » (the processing delay necessary to generate and deliver information and the reporting interval used in the system) et “timeliness” (the availability of information for decision making).	Fox et al. (1994)	
	Currentness	A value at a given time : value can be up-to-date or out-of-date	Iphar et al. (2015)	

	Timeliness	The degree to which data represents reality at the required point in time.	Wei (2016)	
	Currency Timeliness	Datasets selected for processing are the latest Measure if the data is acquired in/evaluated during/adequate for the stated time period	Marev et al (2018)	
	Fraîcheur	Ensemble des facteurs qui capturent le caractère récent et le caractère d'actualité d'une donnée entre l'instant où elle a été extraite ou créée dans la base et l'instant où elle est présentée à l'utilisateur	Berti-Equille (2004)	
	Currency Timeliness	Currency measures how promptly data are updated Timeliness measures how current data are, relative to a specific task.	Scannapieco et al. (2005)	
	"Outdated state"/Currentness	All instances that represent an outdated state in comparison to its data source and thereby approximate timeliness assuming the data source has the closest data representation compared to its real-world state	Fürber & Hepp (2011)	
	Timeliness	An important aspect of data is their update overtime	Batini et al. (2009)	

		Remarque : les auteurs ne définissent pas directement cette dimension mais « dilue » celle-ci dans les notions de « Currency et Volatility »		
	Timeliness Punctuality	Refers to the period between the availability of the information and the event or phenomenon it describes Refers to the delay between the date of the release of the data and the target date (the date by which the data should have been delivered)	European Parliement (2009)	
Consistency **Cohérence**	Consistency	The absence of difference, when comparing two or more representations of a thing against a definition	DAMA (2013)	16
	Comparability	The measurement of the impact of differences in applied measurement tools and procedures where data are compared between geographical areas, sectoral domains, or over time Remarque : les auteurs ont choisi d'agrégé cette dimension avec la dimension de « *facilité d'utilisation ou de manipulation* »	Stróżyna et al. (2007)	
	Consistency	Data consistency refers to whether the logical relationship between correlated data is correct and complete. It usually means that the same data that are located in different storage areas should be considered to be equivalent	Cai & Zhu (2015)	

	Consistency	Implies that not two or more values conflict with each other.	Pawar (2017)	
	Consistency	A dataset should be free of contradictions, while consistency at the data level focuses on the degree to which the format and the value of the data conform to the predefined schema of a given dataset.	Behkamal et al. (2014)	
	Consistency	Data is said to be consistent with the respect to a set of data model constraints if it satisfies all the constraints in the set.	Fox et al. (1994)	
	Consistency	Coherence of information within a message or between messages	Iphar et al. (2015)	
	Cohérence	Quantité de valeurs satisfaisant l'ensemble des contraintes ou règles de gestion définies	Berti-Equille (2004)	
	Consistency	The consistency dimension captures the violation of semantic rules defined over (a set of) data items (Intra-relation integrity constraints)	Scannapieco et al. (2005)	
	Consistency and Synchronization	A measure of the equivalence of information stored or used in various data stores, applications, and systems, and the processes for making data equivalent.	McGilvray (2008)	
	Consistency	The consistency dimension refers to the violation of semantic rules defined over a set of data items.	Batini et al. (2009)	

	Consistency	Consistency refers to data values in one data set being consistent with values in another data set	Loshin (2006)	
	Representational consistency :	The extent to which data are always presented in the same format and are compatible with previous data	Wang et Strong (1996)	
	Consistency	Similarity when comparing two or more representations of something against its definition.	Wei (2016)	
	Consistency	Non-adherence to the proscribed pattern/standard	Gitzel et al. (2015)	
	Consitent representation	The extent to which data are always presented in the same format	Pipino et al (2002)	
Accuracy **Précision**	Accuracy	The degree to which data correctly describes the “real world” object or event being described	DAMA (2013)	16
	Accuracy	The closeness of estimates to the unknown true values <u>Remarque :</u> les auteurs ont choisi d’agréger cette dimension avec la dimension de « *complétude (completeness)*»	Stróżyna et al. (2007)	
	Accuracy	To ascertain the accuracy of a given data value, it is compared to a known reference value	Cai & Zhu (2015)	

	Accuracy	Definition of standard values, for instance in the case of the speed of vessels	Iphar & al. (2015)	
	Accuracy	Data should be as accurate as possible thus a strict requirement for correct quality measurements	Marev et al (2018)	
	Accuracy	Data accuracy refers to the degree with which data correctly represents the “real-life” objects they are intended to model. In many cases, accuracy is measured by how the values agree with an identified source of correct information (such as reference data)	Loshin (2006)	
	Accuracy	The degree to which data correctly describes the ‘real world’ object or event being described	Wei (2016)	
	Accuracy	A measure of the correctness of the content of the data (which requires an authoritative source of reference to be identified and accessible).	McGilvray (2008)	
	Accuracy	Accuracy measures the distance between a value v and a value v’ which is considered correct. Remarque : les auteurs précisent que la dimension accuracy est de deux type. « Syntatic accuracy » (measured by means of comparison functions that evaluate the distance between v	Scannapieco et al. (2005)	

		and v'.) et "semantic accuracy" (captures the cases in which v is a syntactically correct value, but it is different from v'.)		
	Accuracy	Syntatic accuracy: defined as the structural validity of a dataset (entities) as well as the appropriateness of the properties which are used for describing the entities. Semantic accuracy: relates to the correctness of a data value in comparison to the actual real world value	Behkamal et al. (2014)	
	Accuracy	A value is syntactically accurate, when it is part of a legal value set for the represented domain or it does not violate syntactical rules defined for the domain	Fürber & Hepp (2011)	
	Accuracy	Refers to the closeness of estimates to the unknown true values	European Parliament (2009)	
	Accuracy	Syntatic accuracy	Batini et al (2009)	
	Accuracy	The extent to which data are correct, reliable, and certified free of error	Wang et Strong (1996)	
	Accuracy	Percentage of accurate cells: Indicates the percentage cells in a dataset that has correct values according to the domain and the type of information of the dataset.	Vetro et al. (2016)	

	Accuracy	Accuracy of a data refers to the degree of closeness of its value *v* to some value *v'* in the attribute domain considered correct for the entity *e* and the attribute *a.* If the data's value *v* is the same as the correct value *v'*, the data is said to be *accurate* or *correct*.	Fox et al. (1994)	
Precision **Degré de précision/détails** **Granularité**	Precision	Precision refers to the measurement or classification detail used in specifying an attribute's domain.	Fox et al. (1994)	3
	Precision	It is a measurement of the degree of detail of the classification of possible values for data (i.e unit level of measurement → inches instead of feet)	Iphar et al. (2015)	
	Level of detail	Richness of information	Gitzel et al. (2015)	
Validity **Validité/Conformité (syntaxe, structure)**	Validity	Data are valid if it conforms to the syntax (format, type, range) of its definition	DAMA (2013)	10
	Integrity	Data with "integrity" are said to have a complete structure. Data values are standardized according to a data model and/or data type. All characteristics of the data must be correct – including business rules, relations, dates, definitions, etc	Cai & Zhu (2015)	
	Validity	Input data is valid in its purposed used.	Pawar (2017)	

	Data Integrity Fundamentals	A measure of the existence, validity, structure, content, and other basic characteristics of the data.	McGilvray (2008)	
	Validity	Conformity of data's syntax (format, type, range) to its definition.	Wei (2016)	
	Compliance	Percentage of standardized columns: Indicates the percentage of standardized columns in a dataset. It just considers the columns that represent some kind of information that has standards associated with it (i.e. geographic information).	Vetro et al. (2016)	
	Consistency	A dataset should be free of contradictions, while consistency at the data level focuses on the degree to which the format and the value of the data conform to the predefined schema of a given dataset.	Behkamal et al. (2014)	
	Consistency	The data consistently follows a set of predefined rules (e.g., format, type, structure).	Marev et al. (2018)	
	Reliability	General coherence of messages with respect to standards and recommendations	Iphar et al. (2015)	
	Conformance	This dimension refers to whether instances of data are either store, exchanged, or presented in a format that is consistent	Loshin (2016)	

		with the domain of values, as well as consistent with other similar attribute values.		
Relevancy **Pertinence**	Relevancy	The degree to which data meet the current and potential needs of the users	Stróżyna et al. (2007)	7
	Relevancy	The extent to which data is applicable and helpful for the task at hand.	Pipino et al. (2002)	
	Relevancy	The extracted information is helpful for the task. Non relevant data should not be considered.	Pawar (2017)	
	Relevancy:	The extent to which data are applicable and helpful for the task at hand.	Wang et Strong (1996)	
	Perception, Relevance, and Trust	A measure of the perception of and confidence in the data quality; the importance, value, and relevance of the data to business needs.	McGilvray (2008)	
	Relevance	Refers to the degree to which statistics meet current and potential needs of the users	European Parliement (2009)	
	Fitness	The amount of accessed data used by users and the degree to which the data produced matches users' needs in the aspects of indicator definition, elements, classification, etc.	Cai & Zhu (2015)	

Accessibility **Accessibilité**	Accessibility and clarity	the conditions and modalities by which users can obtain, use and interpret data	Stróżyna et al. (2007)	6
	Accessibility	Accessibility refers to the difficulty level for users to obtain data Remarque: les auteurs décrivent également la dimension « *Authorization* » (refers to whether an individual or organization has the right to use the data) qui peut être considérée comme partie intégrante de l'accessibilité.	Cai & Zhu (2015)	
	Accessibility	The extent to which data are available or easily and quickly retrievable.	Wang et Strong (1996)	
	Accessibility	Data must be delivered at the right time, being easily and quickly accessible.	Marev et al. (2018)	
	Accessibility & Clarity	Refer to the conditions and modalities by which users can obtain, use and interpret data	European Parliement (2009)	
	Accessibility	The extent to which data are available or easily and quickly retrievable.	Pipino et al (2002)	
Ease of use/manipulation	Coherence	The adequacy of the data to be reliably combined in different ways and for various uses	Stróżyna et al. (2007)	6

Facilité d'utilisation et/ou de manipulation		Remarque : les auteurs ont choisi d'agréger cette dimension avec la dimension de « *Consitency (cohérence)* »		
	Ease of operation Flexibility	The extent to which data are easily managed and manipulated (i.e., updated, moved, aggregated, reproduced, customized) The extent to which data are expandable, adaptable, and easily applied to other needs	Wang et Strong (1996)	
	Structure	Structure refers to the level of difficulty in transforming semi-structured or unstructured data to structured data through technology.	Cai & Zhu (2015)	
	Ease of use and Maintainability	A measure of the degree to which data can be accessed and used, and the degree to which data can be updated, maintained, and managed.	McGilvray (2008)	
	Coherence	Refers to the adequacy of the data to be reliably combined in different ways and for various uses	European Parliement (2009)	
	Ease of manipulation	The extent to which data is easy to manipulate and apply to different task	Pipino et al (2002)	
Interpretability **Interprétable**	Definition/Documentation	Definition/document consists of data specification, which includes data name, definition, ranges of valid values, standard formats, business rules, etc.	Cai & Zhu (2015)	4

	Interpretability Ease of understanding	The extent to which data are in appropriate language and units and the data definitions are clear. The extent to which data are clear without ambiguity and easily comprehended	Wang et Strong (1996)	
	Interpretability Understandability	The extent to which data are in appropriate language, symbols and units and the data definitions are clear. The extent to which data is easily comprehended	Pipino et al. (2002)	
	Understandability	Indicates the percentage of columns in a dataset that is represented in a format that can be easily understood by the users and it is also machine-readable	Vetro et al (2016)	
MetaData **Métadonnées/Spécification des données**	MetaData	Data producers need to provide metadata describing different aspects of the datasets to reduce the problems caused by misunderstanding or inconsistencies.	Cai & Zhu (2015)	3
	Traceability Compliance	eGMS Compliance: Indicates the degree to which a data set follows the e-GMS standard (as far as the basic elements are concerned, it essentially boils down to a specification of which Dublin Core metadata should be supplied) Five star Open Data: Indicates the level of the 5 star Open Data model in which the data set is and the advantage offered by this reason	Vetro et al (2016)	

	Percentage of columns with metadata	Indicates the percentage of columns in a dataset that has associated descriptive metadata. This metadata is important because it allows to easily understanding the information of the data and the way it is represented.		
	Data Specifications	A measure of the existence, completeness, quality, and documentation of data standards, data models, business rules, metadata, and reference data	McGilvray (2008)	
Credibility/Believability **Crédibilité/Confiance**	Credibility	It refers to the objective and subjective components of the believability of a source or message	Cai & Zhu (2015)	5
	Believability Traceability Reputation	The extent to which data are accepted or regarded as true, real, and credible The extent to which data are well documented, verifiable, and easily attributed to a source. The extent to which data are trusted or highly regarded in terms of their source or content	Wang et Strong (1996)	
	Believability Reputation	The extent to which data are accepted or regarded as true and credible The extent to which data are trusted or highly regarded in terms of its source or content	Pipino et al. (2002)	

	Integrity	Enhancement of consistency with temporal recording of actions	Iphar et al. (2015)	
	Believability	The extracted data is valid and credible	Pawar (2017)	
Security **Sécurité**	Integrity	Data cannot be modified in an unauthorized or undetected manner	Cai & Zhu (2015)	3
	Access security:	The extent to which access to data can be restricted and hence kept secure.	Wang et Strong (1996)	
	Security	The extent to which access to data is restricted appropriately to maintain its security.	Pipino et al. (2002)	
Informative value **Expressivité**	Auditability Readability	Auditability means that auditors can fairly evaluate data accuracy and integrity within rational time and manpower limits during the data use phase. The ability of data content to be correctly explained according to known or well-defined terms, attributes, units, codes, abbreviations, or other information.	Cai & Zhu (2015)	3
	Comparability	Refers to the measurement of the impact of differences in applied statistical concepts, measurement tools and procedures where statistics are compared between geographical areas, sectoral domains or over time	European Parliement (2009)	

	Presentation Quality	A measure of how information is presented to and collected from those who utilize it. Format and appearance support appropriate use of the information.	McGilvray (2008)	
Volatility **Volatilité**	Volatility	Volatility measures the frequency according to which data vary in time (une donnée qui ne varie pas, comme une date de naissance, a une volatilité de 0)	Scannapieco et al. (2005)	2
	Data Decay	A measure of the rate of negative change to the data. volatile data requiring a high reliability level require more frequent updates than data with a lower decay rate.	McGilvray (2008)	
Value-added **A valeur ajoutée**	Value-added	The extent to which data are beneficial and provide advantages from their use.	Wang et Strong (1996)	3
	Transactability	A measure of the degree to which data will produce the desired business transaction or outcome.	McGilvray (2008)	
	Value-added	The extent to which data is beneficial and provides advantages from its use.	Pipino et al. (2002)	
Objectivity **Objectivité**	Objectivity	The extent to which data are unbiased (unprejudiced) and impartial.	Wang et Strong (1996)	2
	Objectivity	The extent to which data are unbiased, unprejudiced, and impartial.	Pipino et al. (2002)	

Cost-effectiveness **Rapport coût-efficacité**	Cost-effectiveness	The extent to which the cost of collecting appropriate data is reasonable.	Wang et Strong (1996)	1
Variety of data and data sources **Variété des sources**	Variety of data and data sources	The extent to which data are available from several differing data sources.	Wang et Strong (1996)	1
Concise **Représentation concise**	Concise	The extent to which data are compactly represented without being overwhelming (i.e., brief in presentation, yet complete and to the point)	Wang et Strong (1996)	2
	Concise representation	The extent to which data are compactly represented	Pipino et al. (2002)	
Free of error **Libre d'erreur**	Free-of-error	The extent to which data is correct and reliable	Pipino et al. (2002)	4
	Accuracy	Ensure that the input data is error free	Pawar (2017)	
	Exactitude	Quantité de valeurs correctes et sans erreur	Berti-Equille (2004)	
	Free of error	This data quality issue refers to what might be called syntactical errors as well as logical errors	Gitzel et al (2015)	

Annexe 12 : Tableau croisé des dimensions et références

Tableau 24 : Synthèse des dimensions présentent dans la littérature

	DAMA	Stróżyna	Cai & Zhu	Pawar	Fox	Berti-Equille	McGilvray	Vetro	Fürber & Hepp	Gitzel	Loshin	Wei	Marev	Behkamal	Wang & Strong	Scannapieco	Pipino	Europ. Parlie.	Batini	Iphar.
Timeliness	X	X	X	X	X	X	X	X	X		X	X	X		X	X	X	X	X	X
Completeness	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X
Consistency	X	X	X	X	X	X	X			X	X	X		X	X	X	X		X	X
Accuracy	X	X	X		X		X	X	X		X	X	X	X	X	X		X	X	X
Uniqueness	X				X		X		X		X	X	X	X		X				
Validity	X		X	X			X	X			X	X	X	X						X
Relevancy	X				X	X		X							X		X	X		
Accessibility		X	X										X		X		X	X		
Ease of use		X	X				X								X		X	X		
Metadata			X				X	X												
Interpretability			X					X							X		X			

	DAMA	Stróżyna	Cai & Zhu	Pawar	Fox	Berti-Equille	McGilvray	Vetro	Fürber & Hepp	Gitzel	Loshin	Wei	Marev	Behkamal .	Wang & Strong	Scannapieco	Pipino .	European Parliement	Batini	Iphar.
Free of error				X		X				X							X			
Credibility			X	X											X		X			X
Security			X												X		X			
Informative value			X															X		
Value-added							X								X		X			
Volatility							X									X				
Objectivity															X		X			
Concise															X		X			
Precison					X					X										X
Cost-effectiveness															X					
Variety of data															X					

Annexe 13 : Synthèse des fréquences des dimensions agrégées

Tableau 25 : Tableau des dimensions agrégées de la littérature et leurs fréquences

Dimensions agrégées (22)	Fréquence (143)
Completeness	19
Timeliness	18
Consistency	16
Accuracy	16
Validity	10
Uniqueness	9
Relevancy	7
Accessibility	6
Ease of use	6
Credibility/Believability	5
Interpretability	4
Free of error	4
Metadata	3
Security	3
Informative value	3
Value-added	3
Precison	3
Volatility	2

Objectivity	2
Concise	2
Cost-effectiveness	1
Variety of data	1

Annexe 14 : Questionnaire

Évaluation des dimensions de la qualité des données liées au trafic maritime de matière première

Description

Ce questionnaire s'inscrit dans le cadre d'une recherche de master en Sciences de l'information de la Haute Ecole de Gestion de Genève ayant pour objectif d'évaluer l'impact de l'utilisation de méthodes de fouille de données pour améliorer la qualité de l'information du trafic maritime de matière première.

Dans le cadre de cette recherche, des données sont récoltées par le biais de méthodes de Web Mining (scrapping et crawling) sur des sites internet directement en lien avec le trafic maritime de matière première. Dès lors, la qualité de ces données nécessite une attention particulière. La notion de "qualité des données" fait référence aux données comme des éléments **fiables** d'information dont l'utilisation permet de satisfaire les besoins et les objectifs que **les utilisateurs** poursuivent dans leur activité.

La qualité des données se décompose en dimensions. Une dimension est utilisée pour représenter une caractéristique spécifique des données. Ainsi une dimension sert à mesurer la qualité des données par le biais des caractéristiques qui la compose de manière objective. Par exemple, la dimension "sécurité" mesure le niveau de sécurité de la donnée en termes de la gestion des accès, la confidentialité etc. mais elle ne mesure pas si la donnée est correcte (Accuracy) ou si cette dernière est à jour (Currentness).

Objectifs de ce questionnaire

Dans un premier temps, ce questionnaire vise à formaliser l'état actuel des données à dispositions de l'utilisateur en fonction des dimensions de la qualité des données. Dans un deuxième temps, ce questionnaire vise à évaluer l'importance accordée par l'utilisateur des données par rapport aux dimensions de la qualité des données en lien avec le trafic maritime de matière première. De ce fait, ce questionnaire doit ensuite permettre l'identification des métriques pertinents à l'évaluation de la qualité des données.

Évaluation de l'importance des dimensions

Pour chacune des dimensions, veuillez indiquer **le niveau de satisfaction** (données actuelles) et **d'importance** que vous leur accordez dans votre activité. Plusieurs dimensions peuvent avoir le même niveau d'importance selon votre propre perception et expertise. Pour chaque affirmation, il vous suffit de sélectionner la case correspondante à votre réponse. Un champ, à la fin du questionnaire, est à disposition pour laisser vos remarques.

Il y a 24 questions dans ce questionnaire.

Evaluation de la qualité des données

***A valeur ajoutée (value-added)**

Les données fournissent des informations jusqu' à ce jour non disponibles, inconnues ou inexploitées. Ces dernières apportent une plus-value à votre activité et offre de nouvelles perspectives de travail.

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
A valeur ajoutée (value-added)	○	○	○	○	○	○	○	○	○	○

***Accessibilité (Accessibility)**

Les données proviennent de source(s) facilement et rapidement accessible(s) (par exemple aucune contrainte d'accès, de quantité ou d'attente) et l'utilisateur peut y recourir avec des technologies disponibles ou aisément déployables.

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Accessibilité (Accessibility)	○	○	○	○	○	○	○	○	○	○

***Actualisation (Timeliness/Currentness)**

Les données sont à jour et mise à disposition dans une période de temps (heures, minutes ou secondes) adéquate au regard de l'information du monde réel qu'elles doivent représenter.

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Actualisation (Timeliness/Currentness)	○	○	○	○	○	○	○	○	○	○

***Cohérence (Consistency)**

Les données sont identiques et non-contradictoires aux mêmes données trouvées dans différentes /autres sources d'information. De plus, les données en lien avec une même entité (navire) ne sont pas contradictoires (par exemple un navire ne peut pas être plus large que long, un tanker ne peut pas avoir le statut de navigation "navigation à la voile etc.).

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Cohérence (Consistency)	○	○	○	○	○	○	○	○	○	○

***Validité/Conformité (Validity/Conformity)**

Les données respectent scrupuleusement des normes/formats et/ou des structures syntaxiques prédéfinies (par exemple le numéro MMSI, la latitude etc.).

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Validité/Conformité (Validity/Conformity)	○	○	○	○	○	○	○	○	○	○

***Unicité/Duplication (Uniqueness)**

Les données respectent le principe d'unicité (présentent qu'une seul fois dans la base de données) ou une même donnée ne devrait pas être représentée sous plusieurs syntaxe/valeur (par exemple le nom d'un navire : Petit Navire, petit navire, P'tit Navire etc.)

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Unicité/Duplication (Uniqueness)	○	○	○	○	○	○	○	○	○	○

***Exhaustivité (Completeness)**

Les données fournissent des informations d'ampleur suffisante, d'une étendue et d'une portée qui correspondent à la tâche à accomplir. Toutes les valeurs de la base de données (champs) sont entièrement complétées.

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Exhaustivité (Completeness)	○	○	○	○	○	○	○	○	○	○

***Facilité d'utilisation/manipulation (Ease of use)**

Les données sont facilement utilisables dans le contexte de la tâche/objectif à accomplir mais peuvent également être facilement enrichies, manipulées et entretenues pour de nouvelles tâches.

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Facilité d'utilisation/manipulation (Ease of use)	○	○	○	○	○	○	○	○	○	○

***Interprétable (Interpretability)**

Le sens des données est immédiatement perceptible et ne nécessitent pas d'informations complémentaires pour leurs compréhension (par exemple traduction, unité de mesure, définition, format etc.).

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Interprétable (Interpretability)	○	○	○	○	○	○	○	○	○	○

***Variété des données (Variety of data)**

Les données sont issues de plusieurs sources permettant ainsi la vérification/comparaison des données (Crosschecking).

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Variété des données (Variety of data)	○	○	○	○	○	○	○	○	○	○

***Libre d'erreur (Free-of-error)**

Les données sont exactes et fiables. Elles nécessitent peu ou pas de traitement dans l'accomplissement de l'activité ou avant leur intégration dans une nouvelle base de données utile à votre activité.

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Libre d'erreur (Free-of-error)	○	○	○	○	○	○	○	○	○	○

***Objectivité (Objectivity)**

Les données proviennent de sources de données impartiales et non-biaisées (aucun traitement préalable n'a modifié les données brutes).

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Objectivité (Objectivity)	○	○	○	○	○	○	○	○	○	○

***Crédibilité (Credibility/Believability)**

Les données proviennent de source confirmées, identifiées comme fiables et bien documentées permettant la vérification des informations.

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Crédibilité (Credibility/Believability)	○	○	○	○	○	○	○	○	○	○

***Rapport coût-efficacité (Cost-effectiveness)**

Les coûts des données et leur traitement sont en adéquation avec l'objectif ainsi que les tâches liées à l'activité.

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Rapport coût-efficacité (Cost-effectiveness)	○	○	○	○	○	○	○	○	○	○

***Représentation concise (Concise)**

Les données sont intelligibles et concises sans fournir d'information superflues et inutiles (par exemple la couleur de la coque du navire).

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Représentation concise (Concise)	○	○	○	○	○	○	○	○	○	○

***Securité (Security)**

Les données doivent être protégées et/ou non manipulable et/ou leur accès limité afin d'en assurer la sécurité.

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Securité (Security)	○	○	○	○	○	○	○	○	○	○

***Volatilité (Volatility)**

Les données prennent en compte la volatilité de l'information sur laquelle elle se base (par exemple une donnée qui ne varie pas, comme les dimensions d'un navire, ont une volatilité de 0).

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Volatilité (Volatility)	○	○	○	○	○	○	○	○	○	○

***Expressivité (Informative value)**

La présentation des données permet à l'utilisateur d'appréhender et/ou d'évaluer et/ou de reconstituer facilement l'information qu'elles contiennent.

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Expressivité (Informative value)	○	○	○	○	○	○	○	○	○	○

***Précision/granularité (Precision)**

Les données contiennent un niveau de détails élevé (par exemple le type exact de la cargaison d'un navire).

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Précision/granularité (Precision)	○	○	○	○	○	○	○	○	○	○

***Métadonnées (Metadata)**

Les données sont enrichies/complétées par des métadonnées permettant de réduire les problèmes d'interprétation et/ou d'incohérence tout en fournissant des informations sur les données (par exemple des statistiques sur l'exhaustivité du jeu de données).

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Métadonnées (Metadata)	○	○	○	○	○	○	○	○	○	○

***Exactitude (Accuracy)**

Les données récoltées décrivent au plus proche le phénomène, l'événement ou la situation du monde réel.

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Exactitude (Accuracy)	○	○	○	○	○	○	○	○	○	○

***Pertinence (Relevance)**

Les données récoltées représentent une information utile et pertinente dans le contexte de la tâche à accomplir.

	Satisfaction des données actuelles					Importance des données				
	Pas du tout satisfait	Peu satisfait	Satisfait	Très satisfait	Hautement satisfait	Pas du tout important	Peu important	Important	Très important	Hautement important
Pertinence (Relevance)	○	○	○	○	○	○	○	○	○	○

***Classement des dimensions**

Selon vous, quelles sont les 5 plus importantes dimensions de la qualité dans votre domaine d'activité ? (Merci de faire votre sélection de la plus importante à la moins importante).

Effectuez un double-clic ou glissez/déposez les éléments de la liste de gauche à la liste de droite. L'élément avec le rang le plus élevé est situé le plus haut jusqu'à celui du rang le moins élevé.

Veuillez sélectionner 22 réponses maximum

Vos choix	Votre classement
Timeliness	
Completeness	
Consistency	
Accuracy	
Uniqueness	
Validity	
Relevancy	
Accessibility	
Ease of use	
Metadata	
Interpretability	

Remarque(s) et/ou précision

Annexe 15 : Synthèse des réponses au questionnaire

Tableau 26 : Classement des dimensions selon leur importance

Dimensions	Satisfaction	Importance	Classement des dimensions selon leur importance
Accessibility	Peu satisfait	Hautement important	1
Completeness	Peu satisfait	Hautement important	2
Accuracy	Satisfait	Hautement important	3
Timeliness	Peu satisfait	Hautement important	4
Uniqueness	Pas du tout satisfait	Hautement important	5
Credibility	Peu satisfait	Hautement important	6
Informative value	Très satisfait	Hautement important	7
Validity	Pas du tout satisfait	Hautement important	8
Concise	Peu satisfait	Peu important	9
Free of error	Peu satisfait	Hautement important	10
Value-added	Peu satisfait	Hautement important	11
Consistency	Satisfait	Hautement important	12
Interpretability	Peu satisfait	Très important	13
Variety of data	Satisfait	Hautement important	14
Objectivity	Pas du tout satisfait	Peu important	15
Metadata	Pas du tout satisfait	Hautement important	16
Precision	Pas du tout satisfait	Hautement important	17

Cost-effectiveness	Satisfait	Important	18
Security	Satisfait	Pas du tout important	19
Relevancy	Satisfait	Très important	20
Volatility	Satisfait	Important	21
Ease of use	Peu satisfait	Hautement important	22

Tableau 27 : Dimensions utiles selon la satisfaction et l'importance

Dimensions	Satisfaction	Importance	Classement des dimensions
Accessibility	Peu satisfait	Hautement important	1
Completeness	Peu satisfait	Hautement important	2
Timeliness	Peu satisfait	Hautement important	4
Uniqueness	Pas du tout satisfait	Hautement important	5
Credibility	Peu satisfait	Hautement important	6
Validity	Pas du tout satisfait	Hautement important	8
Free of error	Peu satisfait	Hautement important	10
Value-added	Peu satisfait	Hautement important	11

Interpretability	Peu satisfait	Très important	13
Metadata	Pas du tout satisfait	Hautement important	16
Precision	Pas du tout satisfait	Hautement important	17
Ease of use	Peu satisfait	Hautement important	22

Annexe 16 : Données issues du scraping

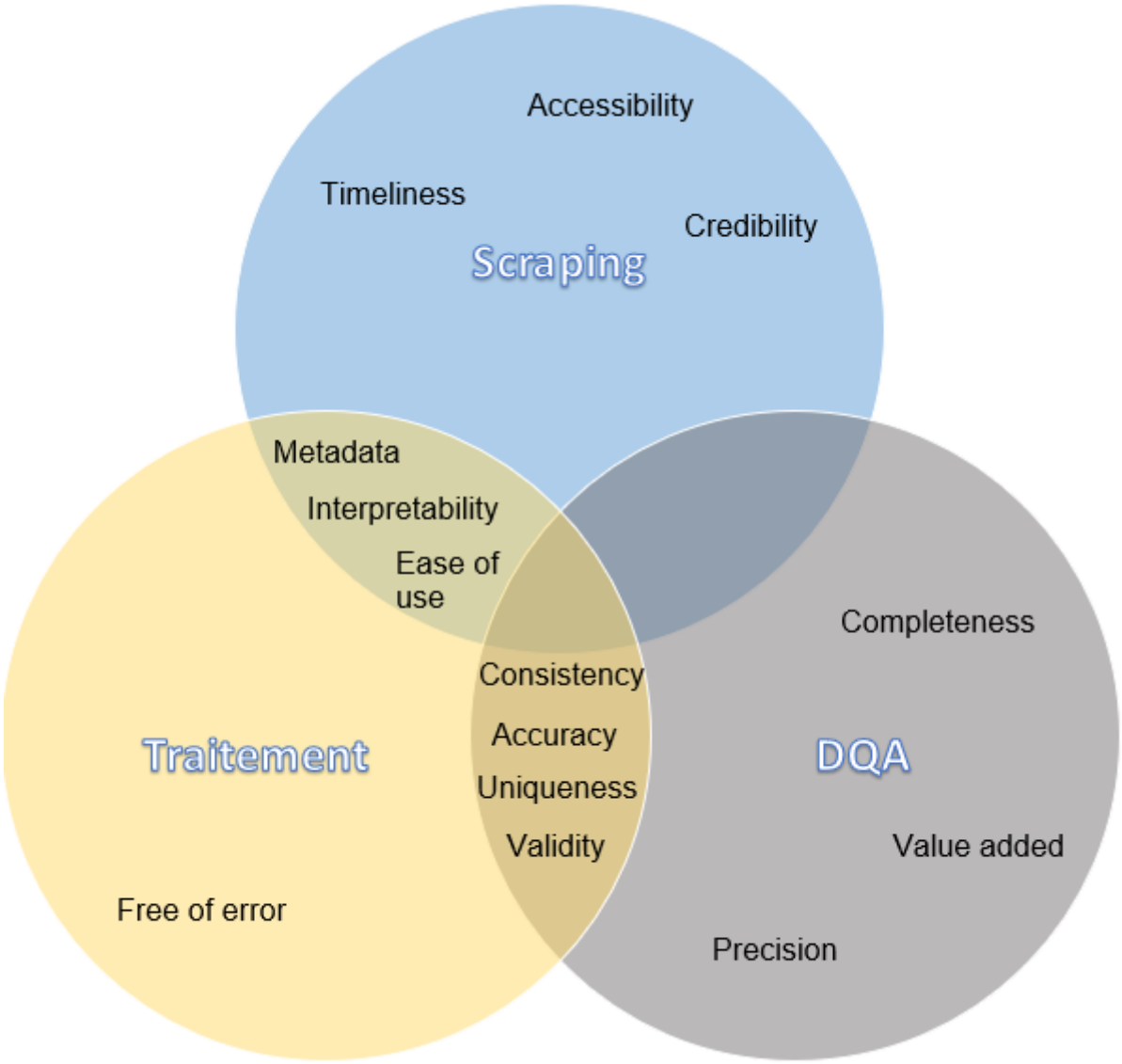
Données d'identification

Port	Nom	Pavillon	IMO	Type
Aarhus (DK)	X			X
Amsterdam (NL)	X	X		
Bordeaux (FR)	X	X		X
Copenhague (DK)	X	X	X	X
Dunkerque (FR)	X			
Fredericia (DK)	X			
Hambourg (DE)	X		X	X
Klaipeda (LT)	X	X	X	
Le Havre (FR)	X	X		X
Niedersachsen (DE)	X	X		X
Rotterdam (NL)	X	X	X	
	11/11	7/11	4/11	6/11

Données de voyage

Port	Statut	Provenance / Destination	ETA/ATA - ETD/ATD	Terminal / Quai	Marchandise	Agent
Aarhus (DK)	X				X	
Amsterdam (NL)	X					X
Bordeaux (FR)	X				X	X
Copenhague (DK)	X	X		X	X	
Dunkerque (FR)	X					X
Fredericia (DK)	X					X
Hambourg (DE)	X	X			X	
Klaipeda (LT)	X	X				X
Le Havre (FR)	X				X	
Niedersachsen (DE)	X			X	X	
Rotterdam (NL)	X	X	X	X		X
	11/11	4/11	1/11	3/11	6/11	6/11

Annexe 17 : Justification des dimensions pour le DQA TMMP

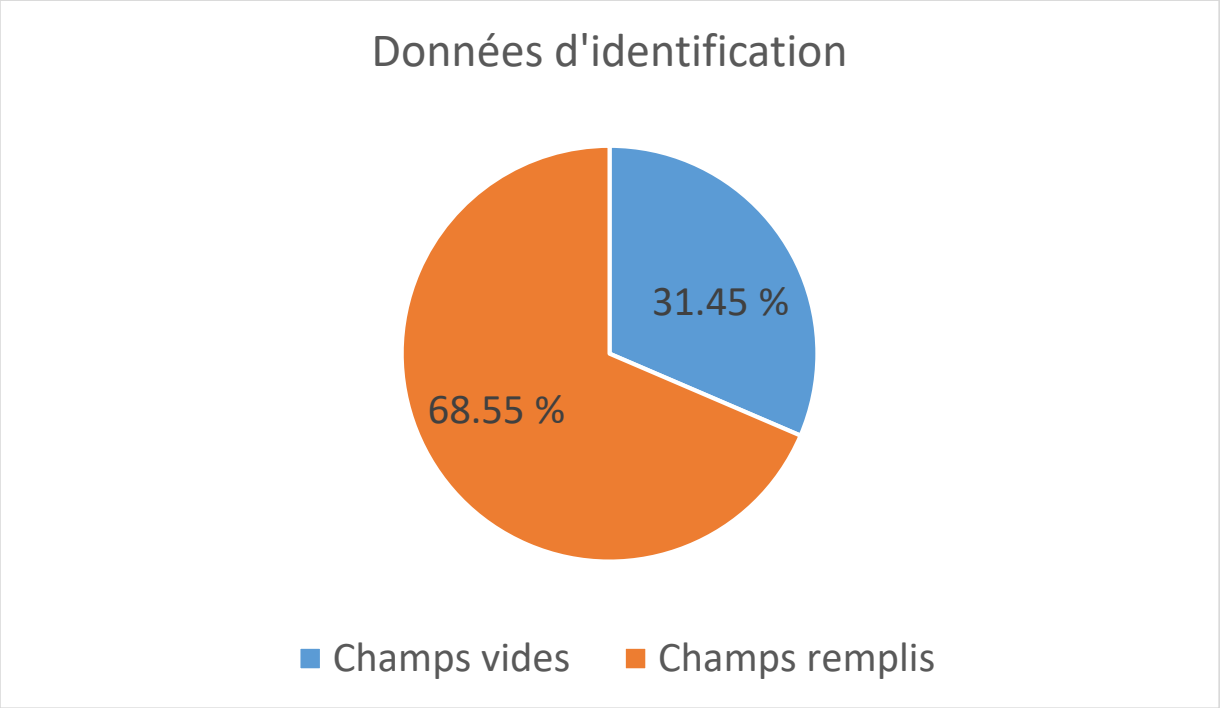


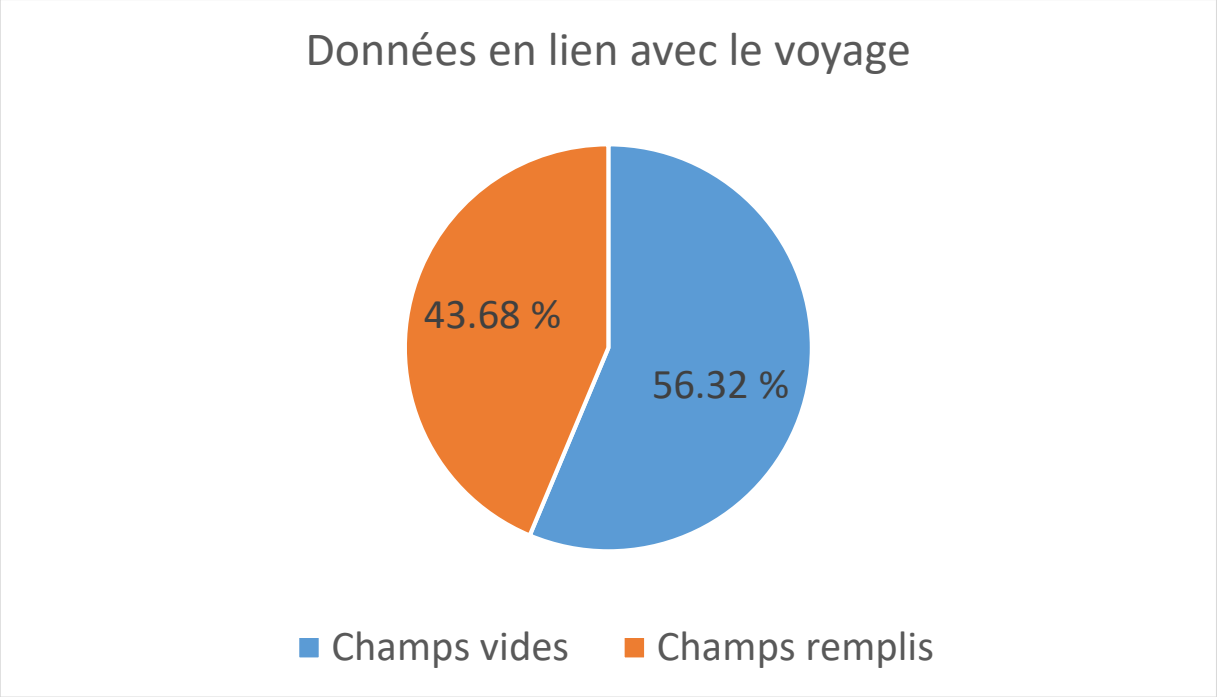
Annexe 18 : Récapitulatifs des données utiles à l'analyse de la complétude

Étiquettes de lignes	Nombre de Nom	Nombre de Pavillon	Nombre de IMO	Nombre de Type
Aarhus (DK)	7			3
Amsterdam (NL)	59	45		
Bordeaux (FR)	2			2
Copenhague (DK)	2	2	2	2
Dunkerque (FR)	13			
Fredericia (DK)	5			
Hambourg (DE)	58		58	58
Klaipeda (LT)	25	25	23	
Le Havre (FR)	44	44	33	44
Niedersachsen (DE)	32	32		32
Rotterdam (NL)	98	98	98	
Total général	**345**	**246**	**214**	**141**

Étiquettes de lignes	Nombre de Statut	Nombre de Provenance	Nombre de Destination	Nombre de ETA	Nombre de ETD	Nombre de Terminal	Nombre de Quai	Nombre de Marchandise chargée	Nombre de Marchandise déchargée	Nombre de Agent
Aarhus (DK)	7			7	7	7	7			
Amsterdam (NL)	59			35	52	22	59			47
Bordeaux (FR)	2	2		2					2	2
Copenhague (DK)	2	2	2	2	2	2	2			
Dunkerque (FR)	13	10	13	10	4		10			13
Fredericia (DK)	5			5	5		5			5
Hambourg (DE)	58			33	25		6			
Klaipeda (LT)	25			22	3		22			25
Le Havre (FR)	44	44	39	44	44	4	44			
Niedersachsen (DE)	32	27	20	32	32	32	32			
Rotterdam (NL)	98			40	53		98			98
Total général	**345**	**85**	**74**	**232**	**227**	**67**	**285**		**2**	**190**

Annexe 19 : Comparaison du taux de remplissage et du taux de vide dans la base de données sur les données d'identification et de voyage





Annexe 20 : Résultats de la validité et de l'exactitude

Tableau 28 : Résultat pour la validité

Variable	Formule	Validité (Validity)
Statut	(345-2) /345*100	99.42
Nom	(345-13) /345*100	96.23
Pavillon	(246-0) /246*100	100
IMO	(214-0) /214*100	100
Type	(141-0) /141*100	100
Provenance	(85-0) /85*100	100
Destination	(74-0) /74*100	100
ETA	(232-0) /232*100	100
ETD	(227-0) /227*100	100
Terminal	(67-0) /67*100	100
Quai	(285-0) /285*100	100
Marchandise déchargée	(2-0) /2*100	100
Agent	(190-0) /190*100	100

Tableau 29 : Résultats pour l'exactitude

Variable	Formule	Exactitude (Accuracy)
Pavillon	(246-0) /246*100	100
IMO	(214-0) /214*100	100
Type	(141-0) /141*100	100
Statut	(345-0) /345*100	100
ETA	(232-0) /232*100	100
ETD	(227-0) /227*100	100
Marchandise déchargée	(2-0) /2*100	100

Annexe 21 : Résultats par valeurs et *poids*

	Statut	Nom	Pavi.	IMO	Type	Prov.	Desti.	ETA	ETD	Terminal	Quai	March. IN	March. OUT	Agent
Completness	100	100	71.30	62.03	40.87	24.64	21.45	67.25	65.80	19.42	82.61	0	0.58	55.07
Accuracy *Pondération*	100 *0.1500*		100 *0.1330*	100 *0.1500*	100 *0.1500*			100 *0.1500*	100 *0.1330*				100 *0.1330*	
Consistency *Pondération*	86.10 *0.3000*							100 *0.300*	100 *0.300*				100 *0.100*	
Validity *Pondération*	99.42 *0.1000*	96.23 *0.0417*	100 *0.0417*	100 *0.1500*	100 *0.1000*	100 *0.0417*	100 *0.0417*	100 *0.1500*	100 *0.1500*	100 *0.0500*	100 *0.0500*		100 *0.0417*	100 *0.0417*
Uniqueness	100													
Precision *Pondération*					43.97 *0.3000*	80.68 *0.2000*		100 *0.1000*	100 *0.100*	100 *0.2000*			100 *0.1000*	
Effective Value added *Pondération*	100 *0.2000*		71.30 *0.0500*			24.64 *0.1000*			65.80 *0.2000*	19.42 *0.1500*	82.61 *0.1500*	0 *0.0500*	0.58 *0.0500*	55.07 *0.0500*
Control Value added *Pondération*		100 *0.2000*		62.03 *0.2000*	40.87 *0.2000*		21.45 *0.2000*	67.25 0.2000						