Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019Mumbai 26-31 May 2019

Critical data quality elements in ports: The case for port rankings

Krenar Ibrahimi[a]*, Silvia Ibrahimi[b]

[a]*Independent researcher, Avni Rustemi street 21, Vlore 9402, Albania*
[b]*Master student at the University of Tirana, Zogu I Bulevard, Tirane 1001, Albania*

Abstract

Port authorities record, collect, and present data about their ports since decades. A complete standardization of port statistics has not been suggested due to diverse port statistical requirements owing to various legal, administrative, and organizational aspects in which information and statistical systems are established and operate, though UNCTAD advised at least since 1971 uniform methods and forms in the collection and presentation of port statistics, introducing also data adequacy and accuracy, availability of accurate and up-to-date information. While research is quite missing in port data quality, this paper focuses port data quality requirements for the special case of port throughput-based rankings of two consecutive years, 2014-2015. First, it will focus on the purposes of port information and the importance and role data quality in general. Ten critical data quality elements will be selected by an extended literature review based on quality data perspectives, framework of topics and methods, and dimensions, principles, and related problems: completeness, availability and accessibility, currency, accuracy, validity, reliability and credibility, consistency, and usability and interpretability. After some specifications of port data and port throughput, these critical data quality elements will be defined and related to other data quality elements, before being analyzed and exemplified on port and container port values of annual throughput used in various worldwide port rankings published by notable organizations, port authorities, and scholars, finding several problems and shortcomings. Conclusions and directions for future research will close this paper that should also be considered as a primer effort to be followed by other scientists in finding new examples of port data and related quality elements, including both data quality dimensions and related problems.

Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY.

Keywords: Port throughput; port ranking; DQ; DQ dimensions; critical DQ elements; port data administration; port data source and comparison; completeness, availability & accessibility, currency, accuracy, validity, usability & interpretability; institutional environment; statistic processes.

* Corresponding author. Tel.: +355-(0)68-526-7-527; fax.
*E-mail address:*krenaribrahimi@gmail.com

Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY

1. Introduction

Since 1971, in its seminal report of 'Port statistics', UNCTAD advised port authorities on "what data should be recorded and how to collect and present them." Furthermore, "it has not seemed advisable to suggest the complete standardization of port statistics, because it is acknowledged that all statistical requirements of all ports are not identical, owing to the great variety of legal, administrative, and organizational circumstances in which information and statistical systems have to be established and operated. Nevertheless ... uniformity is needed in the methods and forms of collection and presentation of port statistics". To the knowledge of this paper's Authors, data presentation uniformity, data adequacy and accuracy, availability of accurate and up-to-date information are the first data quality dimensions used in ports, though not under this label, but since there is no research on the information quality of port data collected, consumed, produced by data processing, and presented.

The development and competition of ports and container ports worldwide have necessitated since decades their yearly ranking in a regional, continental, or global scale, to indicate the extent of changing trade and port related challenges. This ranking is made by globally renowned institutions like the International Association of Ports and Harbors (IAPH), American Association of Port Authorities (AAPA), Shanghai International Shipping Institute (SISI), port authorities like Port of Rotterdam, Port of Antwerp, etc. and scholars interested in port data research. A close observation and analysis of ranking ports and container ports worldwide revealed shortcomings related to some key port data elements. These findings motivated this paper aim that is to relate port data errors and problems to quality elements, assisting port authorities (PAs) in improving their data quality management.

This paper first underlines the importance and role of information as both knowledge and data and of data quality. Then, it makes a review and selection of data quality elements in general and finally tries to identify similar elements and related errors or problems in the existing port and container port rankings. By this doing, this paper tries also to stimulate other researchers to move on this direction in order to increase port data quality and diminish the database of errors that should stay out of port rankings.

Next section will focus on the importance and role of information as both knowledge and data, including also the importance and role of port information and data quality. Section 3 will review the data quality elements in general, while section 4 will identify them and related problems in port ranking data, which will be after a short introduction on the meaning and composing elements of port throughput. Conclusions and further research will close the paper.

2. On the importance and role of port information and port data

The terms data and information are often used synonymously (Wang 1998). The term information, especially as asymmetric information has been cited two times in the motivations of 5 Nobel laureates in economics: Mirrlees and Vickrey in 1996 and Akerlof, Spence and Stiglitz in 2001. Stigler (1961), another Nobel laureate, wrote a paper 'The economics of information' arguing that "One should hardly have to tell academicians that information is a valuable resource: knowledge is power", adding also that the quality of goods, including the quality of information, has not been specified by economics. Like Stigler, Stiglitz (2008) argued that he realized, as he was beginning his work on the economics of information, "that knowledge and information are very similar", admitting later that "economics of [asymmetric] information constitutes a revolution in economics, upsetting longstanding presumptions, including the presumption of market efficiency" (Stiglitz 2016). Asymmetric or imperfect information affects markets up to deny the existence of their invisible hand (Stiglitz 2016) and "both internal organization of firms and its external relations with labor, capital and product markets ... having important behavioral implications ... The appropriate way to look at the whole set of firm decisions – relating also to employment, production, pricing, investment (including inventory changes) and *research* – is as a dynamic portfolio problem" (Greenwald and Stiglitz 1990).

Mattessich (1993) has analyzed the different meanings of semantic, useful, efficient, or genetic information. He argued that a signal or data (datum) might be regarded as a medium possibly carrying some information; a message might be seen as data, information, or knowledge in the communication process, whose metaphor is transportation; and the quantity as well as the *quality of information* has to be distinguished from the information itself. He recalled the history of the information economics from the pioneering works of Jacob Marschak et al., Stigler and Fritz Machlup, and considered the information economics as an extension of the decision theory, a truly interdisciplinary subject (Hansson 1994) that 'transfers' the term information from economics to the field of management science.

Porter and Millar (1985) showed how the information revolution expressed by IT advances affects competition and sources of competitive advantage and what strategies should pursue a company, based on the strategic concept of 'value chain' which divides a company's activities into technologically and economically distinct activities, called 'value activities', embedded in a larger stream of activities termed 'value system'. Linkages not only connect value activities inside a company but create interdependencies between its value chain and those of its suppliers and channels, being a powerful source of competitive advantage because of the difficulty rivals have in perceiving them and trade-offs resolved across organizational lines.

In the case of ports, the value chain and value activities are not other than the logistics and supply chains in which ports and various port activities are embedded, as Robinson (2002) argued. The argument has been underlined by several port scholars as a crucial one for port future port development, organization, and performance.

2.1 On the purposeful importance of port information

"There are several reasons for collecting statistical and other data related to ports. The traditional one is to show the role of the port within the national economy. This appears in the amount of investment expenditure, the number of ships visiting the port and their tonnages, volumes of goods loaded and discharged, classified by main group of commodities, the number of workers engaged in the port industry, etc. ... Statistical and other data are used as tools for improving port operations. The management may wish to compare, on a continuous basis, the actual port activity with its potential. The data collected for this purpose should provide an intimate understanding of the functioning of the port. This is essential in order that the necessary decisions for increasing the efficiency of the port can be taken. The port activity is a complex one, because most of its components are closely interrelated. As a result, a decision which is good for one sector may produce unfavourable effects in other sectors of the port. A sound decision should therefore be preceded by a detailed analysis of the possible effects on the whole system. This is only possible in the relations between the parts of the system can be quantified with the help of *accurate* data. ...

Another purpose of collecting information and presenting it in a systematic form is to provide an appropriate basis for planning port development. The concern of PAs and national planners with port planning is justified by the very large cost of providing and maintaining the port facilities, ... by the frequent indivisibility of port investments, by the difficulty of changing the layout of a port once made and also by the uncertainty regarding forecasts of future traffic and new technologies. The problem of how much, where and when to invest is thus of crucial importance, because a mistake may have a strong negative influence for a very long time to come. Hence the need for having *adequate* and *accurate* information, since this forms the real basis for any decision-making process.

The first task in this respect is to forecast the flow of goods and the ship traffic for future years, taking into account, among other considerations, the evolution towards new technologies in ship design, handling methods, and types of packaging. This exercise is first based on a close examination of the port traffic during the recent past. This traffic has to be analysed by main types of commodities, in order to identify the existing trends. In some ports, the transit traffic represents a large fraction of the total traffic handled and special attention should be therefore be given to the data needed for forecasting this additional transit traffic.

The second task is to compare individual investment projects in the port in order to determine the optimum project. Here again, the interrelations between all parts of the port should be considered, in order to appraise the over-all result of any individual project. The information and statistical system of the port has to provide all the data, including cost data, necessary to quantify these relationships.

Owing to the complexity of port planning, many ports request the help of external consultants or international agencies having competence to provide technical assistance in the field of ports. This in no way reduces the data requirements; in fact, the value of such external assistance depends to a large extent on the *availability* of *accurate* and *up-to-date* information, both statistical and non-statistical, that the external consultant gains the knowledge of the functioning of the port.

In addition, it should be noted that information and statistical data related to the port may be used in other fields of research. In effect, ports are increasingly regarded as links between various means of transport and many data concerning these means of transport are therefore concentrated in ports. As far as shipping is concerned, it can be said that ports are an important source of data for studies related to subjects such as structure and level of freight rates, national shipping policy, organisation of shipping services, etc. ..." (UNCTAD 1971).

2.2 On the importance and role of data quality

Researchers and practitioners have moved beyond establishing information quality (IQ) as an important field to resolving problems from IQ definition, measurement, analysis, and improvement to tools, methods, and processes (Wang 1998). IQ is critical in organizations (Lee et al. 2002). Many important applications increasingly require access to corporate functional and product databases which have disparate levels of data qualities (Wang and Guarascio [WG] 1991). High-quality data can be a major business asset, a unique source of competitive advantage, while poor quality data can lower customer and employee job satisfactions, leading to excessive turnover and the resulting loss of key process knowledge, as can breed organizational mistrust and make it hard to mount efforts that lead to needed improvements (Herzog 2007). Poor data quality (DQ) can have substantial social and economic impacts (WG 1991, Wang and Strong [WS] 1996) and severe impact on the overall effectiveness of organizations (Wand and Wang [WW] 1996).

Though organizations have increasingly invested in technology and human resources to collect, store, and process vast quantities of data, they often find themselves stymied to translate this data into meaningful insights they can use to improve business processes, make smart decisions, and create strategic advantages. Issues surrounding the quality of data and information that cause these difficulties range in nature from the technical (e.g., integration of data from disparate sources) to the non-technical (e.g., lack of a cohesive strategy across an organization ensuring the right stakeholders have the right information in the right format at the right place and time). Although there has been no consensus about the distinction between data quality and information quality, there is a tendency to use data quality to refer to technical issues and information quality to refer to non-technical issues, argue Zhu et al. (2014) who do not make such distinction and use the term data quality to refer to the full range of issues.

Dr. Genichi Taguchi's quality loss function, associating a loss to a quality characteristic deviated from its target value, can be easily extended to DQ. If the quality levels associated with the data elements used in decision-making activities are not at desired *specifications* (*thresholds*), calculations or decisions made based on this data will not be accurate, resulting in huge losses to the organization. Because of the adverse impacts that poor-quality data can have, organizations increase the focus on business DQ, viewing data as a critical resource like others such as people, capital, raw materials, and facilities, and starting to establish a dedicated data management function in the form of the chief data office (Jugulum 2014).

Haug and Arlbjørn (2010) consider 5 barrier themes that prevent companies from achieving high data quality as a lack of: delegation of responsibilities for maintenance of data, rewards for ensuring valid data, data control routines, employee competencies, and user-friendliness of the software managing data. Organizations do best at improving the data quality of their key database when their top management leads the way and is totally committed to such efforts; nevertheless, individual analysts can use the tools to make substantial quality improvements (Herzog 2007).

The relevance of data quality in both decisional and operational processes is recognized by several [national and] international institutions and organizations (Batini and Scannapieco 2006), some of which releasing guidelines or codes of good practice for the usefulness of (data) quality in statistics like Eurostat (2011), UK Office for National Statistics (2013), Statistics Canada (2009), Australian Bureau of Statistics (2009), etc. For instance, Eurostat's (2011) mission is expressed as: "We provide the European Union and the world with high quality information on the economy and society at the European, national, and regional levels and make the information available to everyone for decision-making purposes, research, and debate."

3. Data quality elements and considerations previously established

DQ is an interdisciplinary field (Zhu et al. 2014). Theoretical views as communication theory and information economics are particularly relevant to data characteristics (WW 1996). Researchers primarily operate in major disciplines as Management Information Systems and Computer Science (Zhu et al.), Statistics (Batini and Scannapieco 2006), etc. Due to its interdisciplinary nature, DQ research covers a wide range of topics and methods (Table 1, a).

Following Garvin's (1987) meaning of high quality that is pleasing the consumer, WG (1991) argued that DQ is defined by the data consumer instead of data producers or managers, such as Information Systems departments. Therefore, researchers and practitioners can use DQ to direct their efforts toward quality data by design for data consumers. Following the quality framework with 8 dimensions (performance, features, reliability, conformance,

durability, serviceability, aesthetics, and perceived quality) identified by Garvin (1984) as a framework for thinking about the basic *elements* of product quality, WG argued that DQ encompasses much more than the simple dimension of *accuracy of data*: A dimension is an underlying construct that data consumers use when evaluating data; a DQ dimension is a set of adjectives or characteristics which most data consumers react to in a fairly consistent way.

Table 1. (a) Topics and methods (Zhu et al. 2014) and (b) dimensions (Wang and Guarascio 1991) of data quality research.

a)

Topics	Methods
1. Data quality impact	1. Action research
1.1 Application area (e.g., CRM, KM, SCM, ERP)	2. Artificial intelligence
1.2 Performance, cost/benefit, operations	3. Case study
1.3 IT management	4. Data mining
1.4 Organization change, processes	5. Design science
1.5 Strategy, policy	6. Econometrics
2. Database related technical solutions for data quality	7. Empirical
2.1 Data integration, data warehouse	8. Experimental
2.2 Enterprise architecture, conceptual modeling	9. Mathematical modeling
2.3 Entity resolution, record linkage, corporate householding	10. Qualitative
	11. Quantitative
2.4 Monitoring, cleansing	12. Statistical analysis
2.5 Lineage, provenance, source tagging	13. System design, implementation
2.6 Uncertainty (e.g., imprecise, fuzzy data)	14. Survey
3. Data quality in the context of computer science and IT	15. Theory and formal proofs
3.1 Measurement, assessment	
3.2 Information systems	
3.3 Networks	
3.4 Privacy	CRM: Customer Relationship Management
3.5 Protocols, standards	KM: Knowledge Management
3.6 Security	SCM: Supply Chain Management
4. Data quality in curation	ERP: Enterprise Resource Management

b)

Complete list of dimensions (DIM), their adjectives, the component loading (CL), and the % of variance (% VAR) explained by the dimension

DIM	ADJECTIVES	CL	% VAR	DIM	ADJECTIVE	CL	% VAR
1	Believability	0.76	1.408	12			1.801
2			1.991		Reputation of Source	0.78	
	Competitive Edge	0.74			Data Reputation	0.73	
	Adds Value	0.72		13			3.079
3			2.867		Same Format	0.70	
	Applicable	0.74			Consistently Represented	0.66	
	Relevant	0.64			Consistently Formatted	0.57	
	Interesting	0.58			Compatable w/Previous Data	0.57	
	Usable	0.53		14			2.676
4			5.361		Cost of Collection	0.83	
	Certified Error Free	0.78			Cost of Accuracy	0.78	
	Error Free	0.78			Cost Effectiveness	0.71	
	Accurate	0.73		15			7.315
	Correct	0.71			Easily Joined	0.75	
	Flawless	0.66			Easily Integrated	0.71	
	Reliable	0.60			Easily Download/Upload	0.67	
	Easy Identification of Errors	0.58			Easily Aggregated	0.65	
	Integrity	0.54			Easily Customized	0.59	
	Precise	0.51			Easily Updated	0.56	
5	Interpretable	0.64	1.881		Easily Changed	0.56	
6			2.911		Manipulable	0.53	
	Easily Understood	0.70			Used for Multiple Purposes	0.53	
	Clear	0.65			Easily Reproduced	0.53	
	Readable	0.56		16	Variety of Data and Sources	0.68	1.449
7			3.971	17			6.544
	Retrievable	0.68			Well-Presented	0.81	
	Accessible	0.66			Form of Presentation	0.72	
	Easily Accessed	0.58			Concise	0.71	
	Speed of Access	0.57			Well-Organized	0.71	
	Available	0.56			Format of Data	0.69	
	Up-To-Date	0.56			Well-Formatted	0.68	
	Easily Retrieved	0.52			Compactly Represented	0.66	
8			1.777		Aesthetically Pleasing	0.63	
	Unbiased	0.76		18			2.741
	Objective	0.71			No Access	0.77	
9	Age of Data	0.58	1.494		Proprietary	0.75	
10			3.451		Access Can Be Restricted	0.63	
	Breadth of Information	0.85			Secure	0.60	
	Depth of Information	0.81		19	Amount of Data	0.75	1.610
	Scope of Information	0.79		20			2.360
11			2.609		Adaptable	0.58	
	Well Documented	0.72			Flexible	0.56	
	Verifiable	0.64			Extendable	0.53	
	Easily Traced	0.56			Expandable	0.51	
					Total % of Variance		59.296

By literature review and brainstorming, WG generated an initial list of 36 DQ attributes (or characteristics) that a first survey expanded to 175, with only 10 attributes mentioned by more than half of participants, thus supporting the use of factor analysis for uncovering the actual underlining quality dimensions. This large list is reduced to 118 variables, eliminating by a pre-test those attributes that 11 respondents did not understand or did not see any relation between the attributes and DQ. The DQ dimensions were uncovered using factor analysis on the 355 viable responses collected from the final second survey: WG chose convergence criteria for stopping the analysis the tolerance level as "the amount of variance an original variable shares with all other variables" of 0.001 and limited the number of principle components using the 'eigenvalue >1' rule. By rotation method, the resulting components consisted of those variables whose rotated component loadings were greater than 0.5, an approach quite rigorous, although it may appear simplistic. From the initial analysis generating 29 principal components which explained 73.09% of the total variance in the data, WG eliminated nine based on additional criteria: importance as rated by respondents and interpretability of the components. WG listed the remaining 20 dimensions (Table 1, b) explaining 59.3% of the total variance as for decreasing component loading: (1) believability, (2) value added, (3) relevancy, (4) accuracy, (5) interpretability, (6) ease of understanding, (7) accessibility, (8) objectivity, (9) timeliness, (10) completeness, (11) traceability, (12) reputation, (13) representational consistency, (14) cost effectiveness, (15) ease of operation, (16) variety of data & data sources, (17) conciseness, (18) access security, (19) appropriate amount of data, (20) flexibility.

The same ordered list of DQ dimensions was attained by WS (1996), first grouping them into 4 categories (Table 2, a) based on a preliminary conceptual framework following Moore and Benbasat (1991), their experiences with data consumers, and Juran's quality concept of fitness for use in terms of design, conformance, availability, safety and field use. WS conducted a two-phase sorting study to confirm that 20 intermediate dimensions indeed belonged to the preliminary categories. The second phase adjusted the original assignment based on the results from the phase 1, eliminating 5 dimensions which were not consistently assigned to any category and ranked highly in terms of importance (traceability, variety of data sources, ease of operation, flexibility, and cost-effectiveness) and moving 1 dimension (completeness) from the accuracy to the relevancy category as it was much more assigned to the latter (see column 3, Table 2, a). The adjustment of dimensions within target categories was confirmed by the high placement ratio (70%), which led WS to refine the four target categories to the extent to which data values: conform

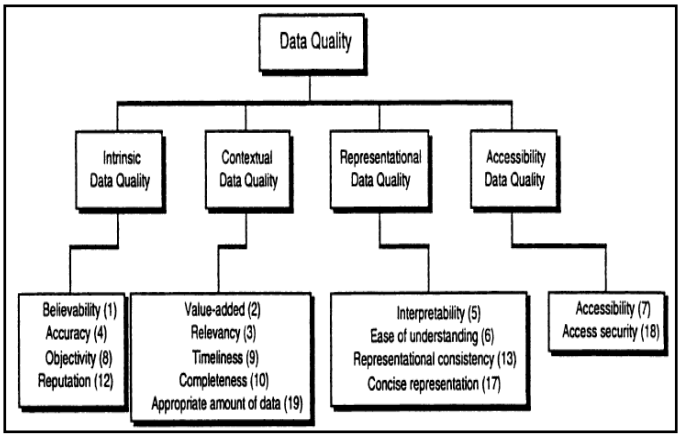
with the actual or true values, are applicable (pertinent) to the task of the data user, are presented in an intelligible and clear way, and are available or obtainable. The results from the phase 2 showed a higher placement ratio (81%) of the 15 remaining dimensions within the appropriate categories. WS concluded that preliminary category labels did not necessarily capture the essence of the underlining dimensions as a group, which were relabeled after reexamination from accuracy to intrinsic DQ and from relevancy to contextual DQ (Table 2, b).

Table 2. A conceptual framework dividing DQ dimensions in four (a) preliminary and (b) adjusted categories (Wang & Strong 1996).

a)

Target category	Dimension	Adjustment
Accuracy of data	Believability	None
	Accuracy	None
	Objectivity	None
	Completeness	Moved to category 2
	Traceability	Eliminated
	Reputation	None
	Variety of Data Sources	Eliminated
Relevancy of data	Value-added	None
	Relevancy	None
	Timeliness	None
	Ease of operation	Eliminated
	Appropriate amount of data	None
	Flexibility	Eliminated
Representation of data	Interpretability	None
	Ease of understanding	None
	Representational consistency	None
	Concise representation	None
Accessibility of data	Accessibility	None
	Cost-effectiveness	Eliminated
	Access security	None

b)



WW (1996) listed in the decreasing level the 26 most cited data quality dimensions (Table 3, a) by counting the appearance in a published article as one citation on a comprehensive literature review. Then, they categorized these dimensions based on the definitions of internal and external views, the latter including interface issues, indicating also whether a dimension is related to the data or to the system (Table 3, b), with timeliness in internal view and reliability in external view appearing to both data- and system-related and timeliness appearing in both internal and external views. But, comparability is appearing nowhere in the dimensional classification shown by WW.

Table 3. (a) 26 cited data quality dimensions divided in (b) internal and external data-/system-related views (Wand and Wang 1996).

a)

Data quality dimension	Citations
Accuracy	25
Reliability	22
Timeliness	19
Relevance	16
Completeness	15
Currency	9
Consistency	8
Flexibility	5
Precision	5
Format	4
Interpretability	4
Content	3
Efficiency	3
Importance	3
Sufficiency	3
Usableness	3
Usefulness	3
Clarity	2
Comparability	2
Conciseness	2
Freedom from bias	2
Informativeness	2
Level of detail	2
Quantitativeness	2
Scope	2
Understandability	2

b)

	Dimensions
Internal View (design, operation)	***Data-related*** accuracy, reliability, timeliness, completeness, currency, consistency, precision ***System-related*** reliability
External View (use, value)	***Data-related*** timeliness, relevance, content, importance, sufficiency, usableness, usefulness, clarity, conciseness, freedom from bias, informativeness, level of detail, quantitativeness, scope, interpretability, understandability ***System-related*** timeliness, flexibility, format, efficiency

Although the diversity of contributions from researchers and practitioners to advance DQ management is valuable, its fundamental aspects relating to DQ dimensions – due to their growing number and evolution and the emergence of new classifications and definitions – have regressed into a level of disparity that does not support a shared understanding of the core knowledge of the discipline (Jayawardene et al. [J+] 2015). J+ used a declarative perspective and a usage perspective focusing respectively on (in)dependent characteristics of data to consolidate a shared understanding of the multiple DQ dimensions having often conflicting interpretations. After reviewing

existing literature and identifying prominent DQ dimension classifications that fit 5 various perspectives – industry practitioners, market leaders of DQ management tools, DQ standards as identified by ISO8000, organizations recognizing DQ importance and developing their own DQ managing frameworks, and academia with rigorous research findings and high-citation level – a rigorous multi-coder approach was applied by J+ to categorize 127 DQ dimensions from 16 sources of thematic analysis in 8 clusters, each possessing dominant quality characteristics (or elements and attributes): completeness, availability and accessibility, currency, accuracy, validity, reliability and credibility, consistency, and usability and interpretability.

The 15 Principles of the code of practice together with the general quality management principles represent a common quality framework in the European Statistical System, grouped in 3 categories, argued Eurostat (2011): (I) Institutional Environment with 6 principles: (1) professional independence, (2) mandate for data collection, (3) adequacy of resources, (4) commitment to quality, (5) statistical confidentiality, and (6) impartiality and objectivity; (II) Statistical Processes with 4 principles: (7) sound methodology, (8) appropriate statistical procedures, (9) non-excessive burden on respondents, and (10) cost effectiveness, and (III) Statistical Output with 5 principles: (11) relevance, (12) accuracy and reliability, (13) timeliness and punctuality, (14) coherence and comparability, and (15) accessibility and clarity.

By focusing on intrinsic aspects of DQ, existing research fails to address the broader concerns of data consumers. While intrinsic DQ aspects are important, organizations also address accessibility and contextual DQ issues, argue Strong et al. (1997) adopting a data-consumer perspective. Their results confirmed the data quality categories and dimensions in WS's previous research, discovering that representational DQ dimensions are underlying causes of accessibility DQ problem patterns. Strong et al. defined a DQ problem as any difficulty encountered along one or more quality dimensions that renders data largely or completely unfit for use and a DQ project as organizational actions taken to address a DQ problem given any recognition of poor DQ by the organization. Wang et al. (1995) used an analogy between manufacturing product and manufacturing data to develop an analysis framework of DQ research as a basis for organizing the DQ literature. The traditional approach of DQ management, referred in literature as ensuring syntactic and semantic correctness, leads to techniques that fail to address important issues to data users, as many databases are plagued with erroneous data or data that do not meet users' needs. Taking a practitioner perspective, Wang et al. devised a framework of seven elements adapted from ISO9000: management responsibilities, operation and assurance costs, research and development, production, distribution, personnel management, and legal function. They found a clear need to develop techniques that help management deliver quality data products, including quality policies and DQ systems, and to study the link between poor data quality and procedures to detect and eliminate problems.

4. Some critical port data quality elements

This section will focus on port ranking by throughput data yearly produced by PAs as original data sources, collected, processed and shared by professional and news agencies from local to global level. For instance, agencies like IAPH, AAPA, SISI, etc., or PAs of major global ports like Port of Rotterdam, Port of Antwerp, etc. produce port rankings. The competitors of PAs, port providers and port users, and port researchers search port data as data consumers, using them for existential and progressing purposes.

Port ranking data, first and foremost, drive decision-making from governmental and port policy level to corporate level of governance in PAs and whatsoever kind of stakeholders for their future investments strategies in infrastructure and to administration and management level of every day's port operating alternatives. The most striking feature of various sources ranking ports by throughput data is that they offer different rankings and values. Data elements refer to a piece of information or any aspect of individual objects taking varying values (Herzog), while critical data elements (CDE) are data critical to success (Jugulum). A standard definition of CDEs typically contains: the element name, a business description and driver, DQ dimensions, source of data, an information subject with which it is associated, stakeholders, privacy criteria, and other key relationships (Jungulum).

In port ranking, CDEs are port throughputs/year which, related to both DQ dimensions (§3) and respective problems as their own main elements intertwined among them, must be considered as port critical data quality elements (PCDQE). The following subsections will first give an overview on port throughput and panorama of port ranking diversity of data and relate them to 10 most significant port QD dimensions, following J+ for the first eight

DQ clusters and Eurostat for the last two, and selected problems considered from 4 perspectives: PAs, port users and port providers as practitioners, port policy makers, port professional agencies, and port researchers.

4.0 On port throughput and throughput-based port rankings

When analyzing port ranking by throughput data, it is important to have a clear understanding on port throughput. In one of its seminal works, 'Berth throughput', UNCTAD (1973) defined berth throughput "as the total number of units which at a given berth or berth area are transferred from one mode of transport to another (of which at least one is craft), during a specified time period," including tonnage that was transferred only from ships alongside a berth. It is obvious to say that a port throughput is the sum total of the throughput in its terminals, which have one or more specialized berths for various types of cargo.

UNCTAD (1971) in 'Port statistics' paid attention to useful data selection in (1) *port facilities and services*, arguing that a port is characterized primarily by its berthing facilities and classifying berths in various types of cargo, including mixed berths, whose number and type depend on the composition of the port traffic and the type and size of ships visiting a port; (2) *ship traffic*; (3) *port operations*; (4) *data on cargo flows and passenger traffic*, which are traditionally established in all ports as they immediately give a picture of the port activity and of changes in the level of this activity from month to month and from year to year; (5) *data on port labour*; (6) *data on costs and revenues*; and (7) *other data*, as ports need other types of information which, although not produced by the port itself, are sometimes important in connection with specific studies; it seems that most important data concern foreign ports, the world merchant fleet, the costs of ships, freight rates, and information on the total trade of the country and projections of the future growth - this information is not of the type which a port itself collects; it is reference material which should be available for use when needed.

"The fourth category of data should be collected in such a way that a detailed analysis of the movements of goods could be undertaken. ... A first step in the classification is ... between cargoes loaded and those discharged. This distinction shows whether or not the inward and outward traffic of a port is balanced, which affects the general pattern of ships' traffic, and hence the port facilities required. A third group concerns transshipment cargoes. Since this operation includes both discharging and loading of the same cargo, it seems logical to record this cargo separately. The cargo may then be classified as to the type of trade. The main types are: (1) Foreign traffic, that is, traffic between ports in two different countries, with the inward goods' movements termed 'exports' and the outward ones termed 'exports', both movements comprising the country foreign trade. Foreign traffic may be subdivided, according to the length of sea voyage involved, into ocean and near-sea traffic; (2) Domestic traffic, that is, traffic between ports in the same country, normally but not always on the same coastline [some ports include cargo movements with river ports]; (3) Transit traffic, that is, traffic physically passing through a port in one country (without entering into that country's foreign trade) having originated in a second country and being consigned to a third country. The transit traffic may leave the country from the same port or from a different port or point, either by sea or by another mode of transport (rail, canal, road, pipeline or air); (4) Entrepôt or re-export traffic, that is, traffic which (a) moves into and out of a port or one of its parts which has been demarcated as a customs-free zone or (b) is imported with the declared intention of being re-exported, usually after minor operations such as packaging, blending, drying, sorting, etc., which leave the goods essentially unchanged. ... Further classification can be made as to the origin and destination of goods. A first breakdown is to classify the goods by country of origin and destination, which in many cases is sufficient to describe the pattern of the port traffic. A more sophisticated classification is by ports of loading and discharge. ... Another classification concerns precisely the types of inland transport used by the port traffic. Distinction should be made between road transport, railway transport, pipeline, inland water transport and coastal ships (feeder services). The purpose of these data is primarily to make possible the best adjustment between the various inland networks (road, rail, inland water) and the port transport network, including the means used to transfer the goods to and from inland transport vehicles. ... None of the classifications suggested ... take into account the nature of goods. From an operational point of view, knowledge of the nature of the goods is less important for the port than knowledge of the type of packaging." UNCTAD (1971).

"Cargo types handled ... determine the physical characteristics of a port and the relevance of various capacity and throughput metrics … different cargo types require different vessels, terminal configurations and handling equipment. Waterborne cargo is generally classified into ... five major types: containerized, dry bulk, liquid bulk, break-bulk,

and Ro/Ro. A large port typically has multiple terminals that together can handle many cargo types; ... individual terminals are usually designed to move a single cargo type." (Bureau of Transportation Statistics [BTS] 2017).

"Port throughput measures reflect the amount of cargo or number of vessels ports handle over time. These measures are affected by many variables beyond physical capacity. ... The throughput statistics included ... are (1) cargo tonnage, (2) container TEU, and (3) vessel calls categorized by commodities carried. It is important to note that the throughput statistics presented ... are annual totals, which can mask seasonal variations in cargo flows ... Cargo tonnage is the most fundamental measure of port and terminal throughput. Cargo tonnage includes the weight of dry bulk and liquid bulk cargo, break-bulk cargo, roll-on/roll-off (Ro/Ro) vehicles and industrial equipment, and the contents of shipping containers. Cargo tonnage does not include the weight of shipping containers themselves, even though movement of empty containers may be a significant part of a port's activity. TEU is a standard measure used throughout the world to measure container movements and the capacity of container ships. While the top 25 [US] ports by TEU are identified by loaded TEU for simplicity since adding empty TEU would not change the list, port throughput statistics presented in the individual port profiles in this report include empty as well as loaded containers to reflect the full volume of activity. USACE does not include foreign empty TEU in its published statistics, so the more complete tabulation of TEU provided by AAPA is used in the port profiles." (BTS).

Though BTS annual report ranks only the top 25 US ports according to port performance throughput-related measures, "to give an indication of the extent of trade growth and the increasing challenges facing [US] ports", it is the unique port document to the knowledge of this paper's Authors including a quality assurance statement: "BTS provides high quality information ... Standards and policies are used to ensure and maximize the quality, *objectivity*, *utility*, and *integrity* of its information" and using these data criteria to select throughput indicators:

- *Availability*. The chosen measures must be readily available for at least the top 25 ports to which they apply (e.g., tonnage for all ports, TEU for container ports, vessel calls and sizes for all ports).
- National *consistency*. The measures must be based on a nationally consistent definition and collection method, and be available for all applicable ports. Ideally, the measures should be available from a single source. If not, multiple sources should be documented and reconciled to ensure reasonable consistency.
- *Timeliness*. The measures should be final and available for the preceding year (a 2015 report includes 2014 data).
- *Relevance* and *clarity*. The measures should be closely connected to the physical activity of ports, terminals, and port infrastructure; and the measures should be understandable to readers who may not be familiar with port or shipping terminology.
- *Accuracy* and *transparency*. The measures should be accurate with acceptable data quality standards and should come from trusted sources.

This paper will focus the throughput-based port rankings produced by the following agencies: IAPH, AAPA, SISI, the Port of Rotterdam (PoR), and Eurostat.

IAPH ranks container ports in various port leagues in a country, regional, continental, and world basis, using as data sources UNCTAD's Review(s) of Maritime Transport, AAPA, ESPO, Japanese government, Containerisation International Yearbook, Manzanillo MIT, and Seaport Alliance Seattle & Tacoma.

AAPA produces world port and container port rankings, limited to top 50 world ports until 2006 and to the yearly lists of top 100 world ports since 2007. AAPA uses as data sources: U.S. Army Corps of Engineers and Waterborne Commerce, AAPA Surveys, various PA internet sites, and Brazil, Mexico, and China agencies, the latter three have replaced the other sources as Shipping Statistics and Containerisation International Yearbooks used until 2006.

SISI produces yearly the 'Global Port development Report' with the main developments in port industry, including the macro situation of global ports in view of the world economy and trade and seaborne transport industry developments, production situation of global ports by analyzing throughput data, trends in port operation and management, business performance and investment tendencies of global port terminal operators, construction of global terminals, (un)loading techniques, technologies and equipment developments, recent progresses in global green ports, SISI-compiled Global Port Development Index offering a comprehensive evaluation of major global ports' development, and future trends and developments forecasted. SISI compares actual and preceding data sourced in PAs' websites and China's Ministry of Transport, which follows the Ministry of Communications as port data provider since China Statistical Yearbook 2008.

The Port of Rotterdam (PoR) in its yearly published 'Port Statistics' [PS] includes mapped data lists of top 20 European and world ports and container ports; renamed as 'Facts & Figures' since 2016, it turned off data and reduced mapped lists to top 10 ports. The Port of Antwerp (PoA) yearly publishes 'Facts & Figures' [FF], including a not-listed map of top 20 global ports. The Port of Rotterdam lists refer to PAs and the Port of Antwerp gives no data sources, as it gives no data.

Eurostat and European Sea Ports Organisation (ESPO) as European agencies provide each year only EU port lists, which are important for comparing port data and port data quality elements with other European port rankings. "Eurostat collect, process and disseminate statistical data for ports. It does this through the cooperation of the National Statistical Institutes of the EU Member States. The methods of collection and processing are different from those used by individual ports. Also the timing of diffusion is slower" (Antonellini 2016). For missing place, only Eurostat lists will be shown, while "It may be noted that, apart from London, data from ESPO is higher than that of Eurostat (or equal for other UK ports)", argued Antonellini, who missed the 14th line corresponding Piraeus port in his Eurostat 2014 ranking of top 20 EU ports.

Tables 4, 5 show top 20 world and European ports and container ports by throughput data (TD) ranked by above cited agencies in 2015. Tables 6, 7 do the same for 2014, making possible comparisons between throughput data.

As shown in Tables 4-7, there are a lot of dissimilarities between agencies for the data they provide for the same ports and their ranking. Quite the same situation, even in port rankings provided by other agencies or ports, is for every year and commodity, as other data for some specific years and ports will appropriately show. Clemenson (2017) uploaded this ranking (in Mdwt) of the top 20 world ports on the port of South Louisiana website: Shanghai, Singapore, Suzhou, Tianjin, Guangzhou, Tangshan, Qingdao, Ningbo, Port Hedland, Rotterdam, Zhoushan, Busan, Rizhao, Yingkou, South Louisiana, Zhangjiagang, Caofeidian, Yosu [aka Yeso] (& Gwangyang), and Yantai. Including in the same world port ranking the port (Suzhou) with one of its member ports (Zhangjiagang) shows how differently institutions and scholars are informed on Chinese ports as Suzhou, integrating Changshu, Zhangjiagang, and Taicang ports (Table 4), and Tangshan consisting of Caofeidian, Jingtang, and Fennan ports areas administrated separately, having different UN/Locodes – both Tangshan and Suzhou ports have also their own UN/Locodes – but considered to be the same port for statistical purposes (Wikipedia).

Table 4. Throughput data (TD) and rankings of top 20 world ports and container ports in 2015 (as for the cited sources).

Rank	**Top 20 world ports** (metric tons)						**Top 20 world container ports** (1000 TEUs)							
	AAPA (1000 tons)		**SISI** (million tons)		**PoR*** (million tons)		**IAPH**		**AAPA**		**SISI**		**PoR***	
	Port	TD	Port	TD	Port	TD	Port	TD	Port	TD	Port	TD	Port	TD
1.	Shanghai	646,514	Ningbo-Zhoushan	889.00	Ningbo & Zhoushan	889.0	Shanghai	36,537	Shanghai	36,516	Shanghai	36,540	Shanghai	36,540
2.	Singapore**	578,846	Shanghai	717.40	Shanghai	717.4	Singapore	30,922	Singapore	30,922	Singapore	30,920	Singapore	30,922
3.	Qingdao	476,216	Singapore	574.90	Singapore	574.9	Shenzhen	24,204	Shenzhen	24,142	Shenzhen	24,210	Shenzhen	24,200
4.	Guangzhou	475,481	Suzhou	543.19	Tianjin	541.0	Ningbo-Zhoushan	20,620	Ningbo	20,636	Ningbo-Zhoushan	20,630	Ningbo & Zhoushan	20,630
5.	Rotterdam	466,363	Tianjin	540.51	Suzhou*****	540.0	Hong Kong	20,073	Hong Kong	20,073	Hong Kong	20,110	Hong Kong	20,100
6.	Port Hedland	452,940	Guangzhou	521.00	Guangzhou	519.9	Busan	19,469	Busan	19,469	Busan	19,430	Busan	19,467
7.	Ningbo	448,828	Qingdao	497.49	Qingdao	500.0	Guangzhou	17,625	Qingdao	17,323	Guangzhou	17,620	Guangzhou	17,590
8.	Tianjin	440,413	Tangshan	490.00	Tangshan	490.0	Qingdao	17,510	Guangzhou	17,097	Qingdao	17,440	Qingdao	17,430
9.	Busan***	347,713	Rotterdam	466.36	Rotterdam	466.4	Dubai	15,592	Dubai Ports	15,585	Dubai	15,590	Dubai Ports	15,590
10.	Dalian	320,658	Hedland	452.94	Port Hedland	452.9	Tianjin	14,090	Tianjin	13,881	Tianjin	14,110	Tianjin	14,110
11.	Kwangyang	272,007	Dalian	415.00	Dalian	415.0	Rotterdam	12,235	Rotterdam	12,235	Rotterdam	12,230	Rotterdam	12,235
12.	Hong Kong	256,488	Rizhao	361.00	Rizhao	361.0	Port Klang	11,887	Port Kelang	11,887	Port Kelang	11,700	Port Klang	11,887
13.	Qinhuangdao	246,550	Busan	359.01	Yingkou	338.5	Kaohsiung	10,264	Kaohsiung	10,264	Kaohsiung	10,260	Kaohsiung	10,260
14.	South Louisiana	235,058	Yingkou	338.00	Busan****	323.7	Antwerp	9,654	Antwerp	9,654	Antwerp	9,650	Antwerp	9,654
15.	Port Kelang	219,786	Southern Louisiana	292.76	South Louisiana	265.6	Dalian	9,450	Dalian	9,591	Dalian	9,450	Dalian	9,450
16.	Houston	218,575	Hong Kong	262.49	Hong Kong	256.6	Xiamen	9,183	Xiamen	9,215	Xiamen	9,180	Xiamen	9,180
17.	Antwerp	208,423	Gwangyang	261.68	Qinhuangdao	253.1	Tanjun Pelepas	9,120	Hamburg	8,821	Tanjung Pelepas	9,100	Tanjung Pelepas	9,130
18.	Xiamen	200,500	Qinhuangdao	253.00	Port Klang****	219.8	Hamburg	8,821	Tanjung Pelepas	8,797	Hamburg	8,970	Hamburg	8,821
19.	Nagoya**	197,947	Yantai	251.00	Shenzhen	217.1	Los Angeles	8,161	Los Angeles	8,160	Los Angeles	8,160	Los Angeles	8,160
20.	Shenzhen	191,037	Shenzhen	217.06	Xiamen	210.0	Laem Chabang	6,780	Long Beach	7,192	Long Beach	7,190	Long Beach	7,190
Total		6,900,343		8,703.79		8,551.9		312,197		311,460		312,490		312,546

* All China's ports include domestic and river trade; ** Freight tons; *** Revenue tons; **** Converted from freight to metric tons by PoR; ***** Suzhou port integrates Changshu, Zhangjiagang and Taicang ports; Highlighted figures have been changed in 2016 for 2015 (changes of a few units are neglected) by SISI.

Next subsections will make a thorough examination of these port rankings and related data, evidencing the main problems through DQ dimensions as clustered by J+ (2015) for the first eight and Eurostat (2011) for the last two.

Table 5. Throughput data and rankings of top 20 European ports and container ports in 2015 (as for the cited sources).

Rank	**Top 20 European ports** (metric tons)						**Top 20 European container ports** (1000 TEUs)					
	AAPA (1000 tons)		**PoR** (million tons)		**Eurostat** (million tons)		**IAPH**		**AAPA**		**PoR**	
	Port	TD	Port	TD	Port	TD	Port	TD	Port	TD	Port	TD
1.	Rotterdam	466,363	Rotterdam	466.4	Rotterdam	436.9	Rotterdam	12,235	Rotterdam	12,235	Rotterdam	12,235
2.	Antwerp	208,423	Antwerp	208.4	Antwerp	190.1	Antwerp	9,654	Antwerp	9,654	Antwerp	9,654
3.	Hamburg	137,824	Hamburg	137.8	Hamburg	120.2	Hamburg	8,821	Hamburg	8,821	Hamburg	8,821
4.	Amsterdam Ports	98,776	Novorossiysk	128.4	Amsterdam	98.8	Bremen*	5,479	Bremen*	5,547	Bremerhaven	5,547
5.	Algeciras - La Linea	91,950	Amsterdam	96.5	Algeciras	79.4	Valencia	4,616	Valencia	4,615	Valencia	4,615
6.	Marseilles	81,920	Algeciras	91.9	Botas	78.1	Algeciras	4,516	Algeciras - La Linea	4,516	Algeciras	4,516
7.	Botas	78,328	Ust-Luga	87.9	Marseilles	77.5	Felixstowe	3,980	Felixstowe	3,676	Felixstowe***	3,984
8.	Bremen*	73,447	Marseilles	81.7	Izmit	64.2	Duisburg	3,600	Gioia Tauro	3,512	Piraeus	3,287
9.	Novorossiysk	73,328	Bremerhaven	73.4	Le Havre	62.9	Piraeus	3,330	Piraeus	3,360	Ambarli/Istanbul**	3,080
10.	Valencia	69,601	Valencia	70.1	Immingham	59.1	Ambarli	3,221	Ambarli	3,062	Marsaxlokk	3,064
11.	Le Havre	68,289	Le Havre	68.3	Valencia	57.6	Marsaxlokk	3,064	Le Havre	2,556	Le Havre	2,559
12.	Izmit (Kocaeli)	64,628	Primorsk	59.6	Bremerhaven	49.8	Le Havre	2,560	Southampton	2,349	Gioia Tauro	2,550
13.	Primorsk	59,606	Grimsby/Immingham**	58.3	Trieste	49.1	Genoa	2,243	Genoa	2,243	Genoa	2,243
14.	Grimsby and Immingham	59,103	Trieste	57.2	Alliaga	48.4	London	2,217	Dublin	2,217	Southampton***	2,108
15.	Trieste	57,161	Constantza	56.3	London	45.4	Barcelona	1,954	Barcelona	1,965	St. Petersburg	1,984
16.	Constantza	56,337	St. Petersburg	51.5	Bergen	43.6	Southampton	1,933	St. Petersburg	1,715	Barcelona	1,965
17.	St. Petersburg	51,513	Genoa	50.2	Genoa	43.4	St. Petersburg	1,715	Zeebrugge	1,569	Zeebrugge	1,569
18.	Genoa	51,299	Dunkirk	46.6	Sines	41.2	Mersin	1,466	Mersin	1,428	Mersin	1,470
19.	Alliaga	48,794	Barcelona	45.9	Piraeus	38.7	Sines	1,332	Sines	1,332	Sines	1,332
20.	Yuzhnyy	48,582	London	45.4	Goteborg	37.8	n.a.	n.a.	La Spezia	1,300	La Spezia	1,300
Total		1,945,272		1,981.8		1,722.2		77,936		77,672		77,883

* Bremen/Bremerhaven; ** Provisional figure; *** Estimated by PoR.

Table 6. Throughput data and rankings of top 20 world ports and container ports in 2014 (as for the cited sources).

Rank	**Top 20 world ports** (metric tons)						**Top 20 world container ports** (1000 TEUs)							
	AAPA (1000 tons)		**SISI** (million tons)		**PoR*** (million tons)		**IAPH**		**AAPA**		**SISI**		**PoR***	
	Port	TD	Port	TD	Port	TD	Port	TD	Port	TD	Port	TD	Port	TD
1.	Shanghai	678,376	Ningbo-Zhoushan	873.47	Ningbo & Zhoushan	873.0	Shanghai	35,285	Shanghai	35,286	Shanghai	35,290	Shanghai	35,290
2.	Singapore**	581,268	Shanghai	755.29	Shanghai	755.3	Singapore	33,869	Singapore	33,869	Singapore	33,869	Singapore	33,869
3.	Guangzhou	500,975	Singapore	581.27	Singapore	580.8	Shenzhen	24,037	Shenzhen	23,798	Shenzhen	23,960	Shenzhen	24,040
4.	Qingdao	465,055	Tianjin	540.00	Tianjin	540.0	Hong Kong	22,226	Hong Kong	22,374	Hong Kong	22,270	Hong Kong	22,200
5.	Port Hedland	446,922	Tangshan	500.80	Tangshan	500.8	Ningbo	18,700	Ningbo	19,450	Ningbo-Zhoushan	19,470	Ningbo & Zhoushan	19,450
6.	Tianjin	445,780	Guangzhou	480.00	Guangzhou	500.4	Busan	18,678	Busan	18,423	Busan	18,678	Busan	18,678
7.	Rotterdam	444,733	Suzhou	479.00	Qingdao	480.0	Guangzhou	16,600	Qingdao	16,624	Guangzhou	16,600	Guangzhou	16,610
8.	Ningbo	429,912	Qingdao	477.00	Rotterdam	444.7	Qingdao	16,580	Guangzhou	16,160	Qingdao	16,580	Qingdao	16,580
9.	Dalian	337,366	Rotterdam	444.73	Dalian	420.0	Dubai	15,249	Dubai Ports	14,750	Dubai	15,250	Dubai Ports	15,200
10.	Busan***	335,411	Dalian	423.00	Port Hedland	372.4	Tianjin	14,061	Tianjin	14,050	Tianjin	14,050	Tianjin	14,060
11.	Hong Kong	297,737	Hedland	421.29	Rizhao	353.0	Rotterdam	12,298	Rotterdam	12,453	Rotterdam	12,298	Rotterdam	12,298
12.	Qinhuangdao	261,702	Busan	346.10	Yingkou	330.7	Port Klang	10,946	Port Kelang	10,736	Port Kelang	10,946	Port Kelang	10,946
13.	South Louisiana	242,578	Rizhao	335.00	Hong Kong	297.7	Kaohsiung	10,593	Kaohsiung	10,593	Kaohsiung	10,593	Kaohsiung	10,593
14.	Port Kelang	217,289	Yingkou	334.00	Qinhuangdao	274.0	Dalian	10,011	Dalian	10,128	Dalian	10,130	Dalian	10,130
15.	Houston	212,561	Hong Kong	295.92	Busan****	266.7	Hamburg	9,730	Hamburg	9,729	Hamburg	9,729	Hamburg	9,279
16.	Nagoya**	207,621	Southern Louisiana	291.83	South Louisiana	264.7	Antwerp	8,978	Antwerp	9,136	Antwerp	8,978	Antwerp	8,978
17.	Antwerp	199,012	Qinhuangdao	274.00	Shenzhen	223.2	Xiamen	8,572	Xiamen	8,572	Xiamen	8,570	Xiamen	8,572
18.	Shenzhen	192,093	Gwangyang	251.15	Xiamen	205.0	Tanjun Pelepas	8,550	Los Angeles	8,340	Los Angeles	8,340	Tanjung Pelepas	8,500
19.	Xiamen	184,604	Yantai	237.00	Antwerp	199.0	Los Angeles	8,340	Tanjung Pelepas	7,897	Tanjung Pelepas	7,600	Los Angeles	8,340
20.	Dampier	172,860	Shenzhen	223.00	Port Klang****	162.0	Long Beach	6,821	Long Beach	6,821	Long Beach	6,821	Jakarta	6,503
Total		6,853,855		8,563.85		8,043.4		310,124		309,189		310,022		310,116

* All China's ports include domestic and river trade; ** Freight tons; *** Revenue tons; **** Converted by PoR from freight to metric tons; Highlighted figures have been changed in 2015 for 2014 (changes of a few units are neglected) by all agencies but AAPA.

Table 7. Throughput data and rankings of top 20 European ports and container ports in 2014 (as for the cited sources).

Rank	**Top 20 European ports** (metric tons)						**Top 20 European container ports** (1000 TEUs)					
	AAPA (1000 tons)		**PoR** (million tons)		**Eurostat** (1000 tons)		**IAPH**		**AAPA**		**PoR**	
	Port	TD	Port	TD	Port	TD	Port	TD	Port	TD	Port	TD
1.	Rotterdam	444,733	Rotterdam	444.7	Rotterdam	421,611	Rotterdam	12,298	Rotterdam	12,453	Rotterdam	12,298
2.	Antwerp	199,012	Antwerp	199.0	Antwerp	180,401	Hamburg	9,730	Hamburg	9,729	Hamburg	9,729
3.	Hamburg	145,673	Hamburg	145.7	Hamburg	126,004	Antwerp	8,978	Antwerp	9,136	Antwerp	8,978
4.	Amsterdam Ports	97,790	Novorossiysk	122.3	Amsterdam	97,098	Bremen**	5,780	Bremen Ports	5,796	Bremerhaven	5,796
5.	Algeciras - La Linea	88,077	Amsterdam	97.8	Algeciras	75,650	Algeciras	4,555	Algeciras - La Linea	4,457	Algeciras	4,555
6.	Marseilles	78,520	Algeciras	95.0	Marseilles	74,426	Valencia	4,328	Valencia	4,442	Valencia	4,442
7.	Bremen Ports	78,236	Marseilles	78.5	Bremen & Bremerhaven	66,442	Felixstowe	4,100	Felixstowe	3,680	Felixstowe	3,700
8.	Novorossiysk	70,000	Bremerhaven	78.3	Le Havre	61,436	Piraeus	3,585	Ambarli	3,445	Ambarli/Istanbul	3,600
9.	Valencia	67,020	Ust-Luga	75.7	Immingham	59,370	Ambarli	3,488	Gioia Tauro	3,062	Piraeus	3,585
10.	Le Havre	66,900	Valencia	67.0	Valencia	55,047	Duisburg	3,400	Marsaxlokk	2,869	Gioia Tauro	2,970
11.	St. Petersburg	61,200	Le Havre	66.9	Trieste	47,265	Gioia Tauro	2,970	Le Havre	2,554	Marsaxlokk	2,900
12.	Grimsby and Immingham	59,400	St. Petersburg	61.2	London	44,489	Marsaxlokk	2,869	St. Petersburg	2,375	Le Havre	2,551
13.	Izmit (Kocaeli)	59,000	Grimsby/Immingham*	59.4	Genoa	43,394	Le Havre	2,550	Genoa	2,173	St. Petersburg	2,382
14.	Trieste	57,200	Trieste	57.0	Piraeus	41,655	St. Petersburg	2,375	Dublin	2,153	Genoa	2,173
15.	Botas	57,000	Constantza	55.6	Barcelona	41,182	Genoa	2,173	Zeebrugge	2,047	Zeebrugge	2,047
16.	Constantza	55,600	Primorsk	53.7	Riga	39,808	London***	2,097	Barcelona	1,894	Barcelona	1,893
17.	Primorsk	53,700	Genoa	51.0	Tees & Hartlepool	39,537	Zeebrugge	2,047	Southampton	1,662	Southampton	1,600
18.	Genoa	51,000	Dunkirk	47.1	Dunkirk	38,919	Barcelona	1,839	Icel (Mersin)	1,484	La Spezia	1,303
19.	Dunkirk	47,100	Barcelona	45.3	Goteborg	36,832	Southampton	1,831	La Spezia	1,301	Sines	1,228
20.	Barcelona	45,000	London	44.5	Southampton	36,688	La Spezia	1,303	Gdansk	1,212	Gdansk	1,212
Total		1,882,161		1,945.7		1,627,254		82,296		77,924		78,942

* Provisional figure; ** Bremen/Bremerhaven; *** Added and taken from 2015 IAPH data for 2014, as London did not figure out in the 2014's list of 19 ports; Highlighted figures, including provisional ones for UK ports, have been changed in 2015 for 2014 (changes of a few units are neglected) by all agencies but AAPA.

4.1 PCDQE1: *Completeness*

J+ recalled many scholarly definitions and characteristics for data completeness, arguing that it is considered in a broad sense and contains several themes: namely, it focuses on handling null values, representing real world objects without omission, and maintaining right volume of data for intended usage can be considered as dominating themes.

Various agencies rank ports in lists of a various number of busiest ports in the world, continent, inter-region (EU, for instance) or country. Only IAPH ranked the list of top 20 European ports missing the 20th port name and data. Also, Antonellini (2016) completely missed a row of data for the 14th port in the 2014 list of Eurostat he showed.

Though the following information is mainly related to the DQ dimension of accuracy (§4.4) and that of reliability and credibility (§4.6), the missed throughput values for various ports instead of others entering in port ranking lists may be also considered as an issue related to the completeness as a DQ dimension. This problem is more frequent in the world port lists for Russian, Chinese, Indian, and Korean ports and in European lists for Russian ports.

Various sources including Russian PAs, institutions or scholars give Novorossiysk as the busiest Russian port at least since 1990, but it has not still been present in AAPA top 50 world ports, as in 2003-4, though its throughput is quite still growing since 2002, its first entry in AAPA with 63,291 ktons. The new growing port of Ust-Luga made its entry in AAPA top 100 world ports only in 2016, though it had an annual throughput higher than 46,786.1 ktons of 2012, as for the statistics of Big Port of St. Petersburg. Also, the port of Vostochny having since 2012 an annual throughput higher than 43 ktons (NCSP) has no place in AAPA lists. The Japanese port of Yokohama is disappeared from AAPA top 100 world ports in 2016, though having had a throughput of 114,743 ktons in 2015, when its PA shows a throughput value of 1.22 times higher – 291,795,408 versus 238,401,976 tons – than usually the busiest Japanese port of Nagoya. The same table furnished by Yokohama PA shows a container throughput of 2,292,517 TEUs, the same value indicated by AAPA that listed Yokohama in its top 100 world container ports in 2016.

The Korean port of Pyeongtaek entered AAPA top 100 world port ranking only in 2016 with a throughput of 112,215 ktons, never being in AAPA lists before 2016, though its port corporation admitted in its December 2015 newsletter that "The port handled more than 100 million tons annually for 3 consecutive years" (GPPC 2015), that is, since 2013. The Chinese port of Nanjing, that Chinese Statistical Yearbook considers as a river port, entered only

in 2008 the AAPA top 100 world port and container port rankings with 111,000 ktons and 1,292,100 TEUs, never before and after, while the other Chinese port of Tangshan disappeared in 2009 like Yokohama in 2016, having had the previous year (2008) the 34th place with 109,000 ktons and entered the top 100 list of AAPA world ports in 2007 with 47,399 ktons. The most striking port data quality element of AAPA rankings is the entry of the US port of Philadelphia at the 83rd place in 2016 top 100 world port ranking with 52,449 ktons, when this port has a throughput value of 22,973,188 tons in the AAPA ranking of top 150 US ports the same year. The same issue in AAPA world port rankings exists for the Chinese ports of Ningbo-Zhoushan, created in 2006 by the combination of Ningbo and Zhoushan ports, and Suzhou, including the Zhangjiagang, Changshu and Taicang ports, having at least since 2006 a total cargo throughput of more than 100 million tons – within 380-540 million dwt in 2011-2015 (Clemenson 2017) – but not yet entered top 100 world ports. Suzhou and Tangshan ports do not appear in China Statistical Yearbook, as Rizhao and Yingkou do since ever and Ningbo-Zhoushan having entered since 2008, but AAPA ranked no one.

AAPA top 100 world ports include 5-6 Indian ports, but since at least 2002 the major port of Kandla in Gujarat, the busiest Indian port since 2008 (various sources, including Indian Ports Association) with an annual throughput of over 45 million tons since 2004-5 and over 70 million tons since 2008 has never been inside. Also, the Sri Lankan port of Colombo entered AAPA rankings in 2016 with 82,221 ktons by not being in AAPA rankings before 2016.

So, ranking directly from nothing to the 32nd place as Nanjing, 37th place as Pyeongtaek, 46th place as Ust-Luga, or 50th place as Colombo port, or disappearing the next year from the 35th place, having more than 114 million tons, as Yokohama port, are very curious movements of ports in AAPA top 100 world ports.

4.2 PCDQE2: *Availability and Accessibility*

J+ links accessibility with *clarity* and *timeliness*, and availability with *ease of use* and *maintainability*, *timeliness* and *punctuality*, *security*, *speed*, and *reliability*, recalling some scholarly definitions and characteristics of both dimensions and their related terms, before arguing that on-time availability and data security perspective are two dominating aspects of this cluster of 2 DQ dimensions. Though timeliness and currency are terms having significant interplay and overlap, there are fundamental differences between timely availability of data versus correct data aging or freshness. Timeliness refers to the time expectation for accessibility and availability of information, a characteristic of getting it when needed. For an efficient database management, timeliness measures the degree to which data can be accessed and used, but also can be updated, maintained, and managed. On the other hand, data accessibility has a security perspective as it can be restricted and hence kept secure, in order to protect information against loss or unauthorized access.

Concerning the timely availability of port data there is much to say about the reasons that several agencies and ports have to not give fresh data in due time (see §4.1). To the actual knowledge of this paper's Authors, Chinese seaports are champions for not delivering data to all port stakeholders as data consumers, including researchers. PoA was self-claimed in its 'FF' "a major international port ... ranking 15th [in 2017, 11th in 2014-16, 10th in 2013, 13th in 2011, and 12th in 2010] in the 20 largest ports in the world". This ranking is related to international maritime freight volumes, as PoA annual reports made clearly known the same years. 'FF 2015' ranks Rotterdam, Singapore, Ningbo Zhoushan, Qingdao, Port Hedland, Shanghai, Tangshan, Rizhao, Tianjin, Nanjing, Antwerp, Gwangyang, Shenzhen, Huanghua, Dampier, Port Klang, Newcastle, Nagoya, Hamburg, and Houston. The PoA website shows that 'FF' are accessible 2 years earlier than its voluminous annual reports. Timely availability of this kind of implicit claims, made explicit only 2 years later, raises questions on integrity and credibility (see §4.5).

The behavior of Chinese or other ports to not deliver port statistics by providing only port data inaccessibility to non-native speakers would be security-based, as well as ESPO strategy that, since 2015, declared to not publish facts and figures for EU ports, but will send them to those wishing to be part of a dedicated mailing list.

4.3 PCDQE3: *Currency*

The currency refers to whether a datum in question is up-to-date and not obsolete and if information is correct despite possible time-related changes, said J+ referring to scholars. So, it is related to the *timeliness* of information, the extent to which data is collected within a reasonable time period and is available within a reasonable timeframe to be used for whatever purpose it is intended; its *aging*, as being appropriate to its use, for the task at hand; its

decay, measuring the rate of negative change to data; its *volatility* as an amount of time it remains valid in activity's context. Therefore, the characteristics of data currency are *timeliness*, acceptable time availability for use from its time of creation, and *freshness*, that means up-to-date and fresh data with respect to its intended use, argued J+.

The question of port data freshness is first related to having data when they are needed, as soon as possible is the best in the perspective of PAs as practitioners, port policy makers and professional agencies for their investment purposes and competitive strategies. This is not a problem for port researchers in the sense that research questions may still involve port data some years later, as we are doing in this paper using port information of several decades.

Annual port throughput data are at best available months after the closure of a fiscal year's activity, a notion in itself a problem in comparing throughput data. Ports in the so-called Anglo-Saxon countries use fiscal years that begins in the middle of a year *y*, say June 1st, ending in the same period of the year *y*+1, say May 31st, while other countries' ports use as fiscal year the calendar year. Eurostat disseminates maritime and port data "according to the periodicity specified [quarterly or annually] (within 10 months after the end of the reference period [quarters or calendar years] for quarterly statistics and within 16 months of the end of the reference year for annual statistics). ... data for a quarter ending in month M are released in month M+10, while annual data for a calendar year ... (month M, year T) are released in March 15 months later (month M+15, year T+2)." (Eurostat website). Antonellini (2016) argued that "diffusion timing is slower, in December we know the data of European ports for the previous year."

From communications with the Port of Barcelona, asking to know which of the different values they give for port throughput 2014 and 2015 values in the documents of two consecutive years, a statistician said: "The traffic data of the Port of Barcelona are data that may vary over time as we depend on the customer's declaration. Sometimes, we receive modifications, extensions or cancellations so in future statistics these changes are reflected. The most recent data is always the closest to reality, so in the case of the statistical data of the throughput 2015 that appear in the report of the year 2016, are the most real and, therefore, more correct." And, this is true for several ports, but there are also other ports that show the same throughput figures anytime in their reports in successive years.

Within the up-to-date logic of port data freshness a lot of comments may regard the information that Tables 4-7 present. For space motive this paragraph will concern only Eurostat's data for top 20 EU ports (Tables 5 and 7). The dissemination policy obliged Eurostat to release port data for both years at least 15 months later than PoR prepared its counterparts, respectively 'PS 2014' in May 2015 and 'PS 2015' in May 2016. However, having collected data from all EU ports, excepting Russian ports in PoR list, which are outside the EU administrative area, Eurostat should have the same figures with them, Rotterdam in particular. Excluding 3 Russian ports, Eurostat list has different throughput data not only for the top 3 European ports, Rotterdam, Antwerp, and Hamburg, but also for the other 12 ports in common with PoR, with lower values excepting London and Immingham having quite the same value in both ranking lists for the year 2014 (Table 7). The same situation is for 2015 (Table 5), where Eurostat list has lower throughput values for 13 European ports in common with PoR list, with Amsterdam and Immingham having higher values: only London has the same value in both lists for the year 2015. So, Eurostat throughput values are even fresher and less accurate than former port data gathered by PoR, which is more interested in its business logic to act quickly and surely towards its competitors. Eurostat, in the policy maker logic of its boss, EU, is decided to go slowly, not having competitive engagements at business level, which is revealed a source of administrative errors by producing poor DQ for doing business, though of a good DQ in respect of their DQ management system.

Port ranking lists showing a unique list of *n* (from 5 to 150) ports and their throughput data ranked during *m* years are utile for presenting data about each ranked port for the given period. But, they are useless for ranking purposes beyond the last or actual year, as a port ranking changes because throughput values of the listed ports change and, most importantly, the names of ports do: some out of *n* ranked ports enter and others disappear in a given year. And, as still occurs, in these lists only the ranking of throughput values for the last or actual year corresponds to the ranking of port names, with all other throughput lists corresponding to precedent years being not ranked. IAPH continues to offer this kind of lists, while PoR since 2016 stopped giving throughputs ranking only port names.

PoR listed ports for three consecutive years up to 'PS 2015'. One may note that Tangshan port, ranked 9th in 2011-12, 7th in 2013, and 5th in 2014 – as 'PS 2012' shows for the period 2010-12 and 'PS 2014' for 2012-14 – has not been in the lists that 'PS 2010' shows for 2008-10 and disappears for the 3-year period 2013-15 that 'PS 2015' shows. On the other side, 'PS 2014' shows the same figures for China ports for 2012 which were provisional in 'Port statistics 2012' and should normally have been rectified according to the real data recorded by Chinese ports. The same may be said for the Grimsby/Immingham port throughput of 2012, which was provisional in 2012 but remained the same

even in 2014, as 'PS 2014' shows; but, 'PS 2015' has rightfully changed for the year 2014 the provisional throughput values of the UK ports that 'PS 2014' shows.

4.4 PCDQE4: *Accuracy*

Accuracy is the first and foremost requirement that many users expect from data (J+). Data accuracy refers to the degree with which data values agree with an identified source of correct information: in this sense, accuracy is related to the process of data creation; the level of accuracy or precision is another aspect driven by consumer needs: data values are correct to the right level of detail; conciseness has a perception component related to user opinion: it is the information to the point, void of unnecessary elements, argued J+ referring to other scholars. Accuracy measures the correctness of the data content, requiring an authoritative source of reference to be identified and accessible; it is the extent to which data are correct reliable and certified free of error. Therefore, verification of accuracy involves comparing the collected data to an external reference source that is known to be valid, capturing data as close as possible to the point of activity. If data accuracy is compromised in any way, then this information should be made known to the data users (J+). Other DQ dimensions are related to accuracy as *verifiability*, the information correctness is verifiable or provable; *validity*, the information is legitimate or valid as for some stable reference source; *reliability*, the data is collected consistently over time and by different organizations, either manually or electronically; and *precision*, attribute values should be precise as per form, linguistics, and granularity (J+). This subsection will apply some of these DQ elements to port data.

Port data in Tables 4-7 essentially have 2 elements asking for accuracy: port names as a string of characters and their respective throughput numerical values. Both elements show problems of format and precision. No standard formats are applied for writing port names: some agencies use characters like &, /, -, etc. in port names including 1-3 words. No standards are applied when writing numerical values: agencies use various decimal places and units.

There is a long list of port names that are variously written by different agencies and even by the same agency: the US port of South Louisiana is also written Southern Louisiana; the Dutch port of Amsterdam is also written Amsterdam Ports; the Chinese port of Ningbo-Zhoushan, today's world busiest port having combined Ningbo and Zhoushan ports in 2006 (China Statistical Yearbook 2009, PoR 'Port statistics 2010'), is written also Ningbo & Zhoushan; the Australian Port Hedland is also written Hedland; Bremen Ports is also written Bremen/Bremerhaven, Bremerhaven, and Bremen Bremerhaven; Dubai Ports is also written Dubai; the Turkish port of Ambarli is also named Ambarli/Instanbul, while the other Turkish port of Mersin is named also Icel (Mersin); the UK port of 'Grimsby & Immingham' is never written so, as for UK Department of Transport, but as Grimsby/Immingham, 'Grimsby and Immingham', or Immingham; the Malaysian port of Tanjung Pelepas is also written Tanjun Pelepas; Gwangyang port of Korea is written Kwangyang; and the busiest Malaysian Port Kelang is written Port Klang: in 2014 lists, PoR ranked Port Klang at top 20 world ports and Port Kelang at top 20 world port containers, having unified both names in Port Klang in its 2015 lists.

Top 20 world ports 2014: AAPA and SISI rank 8 different port names, only 1 equal value (Rotterdam) and 1 close one (Hong Kong); AAPA and PoR rank 4 different port names, 3 equal values (Rotterdam, Hong Kong, Antwerp) and 2 close values (Singapore, Guangzhou); SISI and PoR rank 4 different port names – including the variously written Ningbo-Zhoushan – 6 equal values (Shanghai, Tianjin, Tangshan, Rotterdam, Qinhuangdao, Shenzhen) and 6 close ones (Ningbo-Zhoushan, Singapore, Qingdao, Dalian, Yingkou, Hong Kong). Top 20 world container ports in 2014: only 2 ports are different, including Nigbo vs Ningbo-Zhoushan, Jakarta in PoR and Long Beach in IAPH, AAPA, and SISI, with a few differences in values.

Top 20 world ports in 2015: AAPA and SISI rank 6 different port names, only 2 equal values (Rotterdam, Hedland) and 3 close values (Singapore, Hong Kong, Qinhuangdao); AAPA and PoR rank 5 different port names, 3 equal value rounded (Rotterdam, Hedland, Port Kelang) and 2 close values (Singapore, Hong Kong); and SISI and PoR rank 2 different port names, 12 equal values and 4 close values (Suzhou, Guangzhou, Qingdao, Hong Kong).

Top 20 world container ports in 2015: only 2 ports are different, including Nigbo vs Ningbo-Zhoushan, Laem Chabang in AAPA and Long Beach in the other lists, while a few differences in value exist. Top 20 European ports and container ports in 2014 and 2015: excepting Russian ports that are not present in Eurostat lists, there are some differences in port names and throughput values for the same ports everywhere.

Great differences may be also observed in both AAPA and Lloyd's lists of top 100 world container ports in 2016.

There is a problem of fundamental importance related to port throughput. Many PAs show different values of port throughput within their documents corresponding to consecutive years and even to the same year. For instance, Port of Antwerp shows a throughput of 199,012,082 tons in 'FF 2015 (for 2014)', while its 'Statistical yearbook [SY] 2014', for the same year shows a port throughput of 199,014,454 tons, and 'SY 2015' shows a 2014 port throughput of 199,017,909 tons. The same occurs for the port throughput 2015: 208,419,668 tons in 'SY 2015', 208,420,432 in 'FF 2016' and 208,424,902 tons in 'SY 2016'. Does PoA or others have the same reasons as Port of Barcelona (§4.3)?

4.5 PCDQE5: *Validity*

Data values must conform to specified business rules and derive from a set of specified calculation formulas. Validity of data refers to data collected in accordance with any applicable rules and definitions to enable over time benchmarking among organizations; it pertains to a specified range of valid values, passing all edits for acceptability (J+). Data validity relates to DQ dimensions as *integrity*, not missing important relationship linkages; *conformance* to a format that is consistent with the domain of values and attribute definition; *specifications* that measure the existence, completeness, quality, and documentation of data standards and models, business rules, metadata, and reference data; *consistency* to physical instances of data recorded and represented in a due format; coherence to the internal consistency of data for various organizations and times and to their adequacy to be reliably combined in different ways and for various uses; and *accuracy*, in the sense of denoting the closeness of computations to the exact or true values, argued J+ recalling other scholars.

In addition to port data format discussed in §4.4, port rankings suffer from the known problem of port throughput units. Some PAs use metric tons; others use freight, revenue, short, or long tons. While AAPA unit for Busan (Pusan) port throughput is revenue tons, as all other South Korean ports, PoR asserts that its throughput data, together with Port Klang's, is converted from freight to metric tons. AAPA unit for Singapore port throughput is freight tons, while other agencies say nothing, therefore being supposed to use metric tons instead.

The problem of data integrity as a missing relationship linkage is shown in PoA claimed rank and the real significance of this rank, related to the international maritime freight volume instead of port annual throughput, as explained in §4.2. Top 20 world-port lists rank the port of Antwerp 17th in 2014-15 (AAPA) and 19th in 2014 (PoR).

Some problems of port data validity may be observed through the above Tables. By observing IAPH data tables (2017), the ports of London and Mersin, with 2,097 and 1,499 kTEUs respectively, should be ranked in 2014 instead of the 20th free place and the last listed, La Spezia with only 1,303 kTEUs. Likewise, the port of Ningbo-Zhoushan should have replaced Ningbo in 2014 ranking, as it has been created in 2006 and its throughput of 19,430 kTEUs for 2014 was known together with other data provided for the period 2007-2016 by IAPH itself in 2017.

IAPH listed only 19 out of 20 top container ports for Europe, including Duisburg river port in 2014, and has only 10 equal values with PoR; their lists of top 20 world ports have 11/20 equal values, including values differing by 1 unit. Throughput data for the China's Zhanjiang port entering the top 20 list of 2016 (SISI 2017) is 220.35 million tons for 2015, which is greater than Shenzhen ranked 20th in the list of 2015 with 217.06 million tons.

PoR ranked Jakarta 20th in 2014 for world container ports, while with 2015's figures PoR gives 6,818 kTEUs for Long Beach in 2014, which means that this port should replace the other with only 6,503 kTEUs in 2014. By observing the data for Suzhou port given by PoR in 2015 for the three years 2013-15, it results that this port should have been part of the top 20 world ports since 2013, therefore ejecting from the list the last one, that of Port Klang.

4.6 PCDQE6: *Reliability and Credibility*

The main aspect of this DQ cluster is the assurance of the trustworthiness of data, argued J+. Other DQ dimensions are related to this cluster: *believability*, accepting or regarding data as true, real and credible; quality and security warranties of information sources that (a) guarantee the information provided with remedies for non-compliance, (b) document its certification in its information quality management capabilities to capture, maintain, and deliver quality information, (c) provide objective and verifiable measures of the information quality provided in agreed-upon quality characteristics, (d) guarantee information protection from unauthorized access or modification; *reputation*, trusting or highly regarding data in terms of their source and content; *objectivity* of presenting unbiased and impartial data; *perception* of syntactic and semantic criteria; *traceability* of the visible information background;

verifiability of the information correctness in a particular activity context; *authority*, as a reputation degree of an information object in a given community and culture; *enterprise agreement of data usage* permitting various users to manage their own versions; *data provenance* including information on creating, updating, transcription, abstraction, validation and transforming data ownership; and *credibility* on the amount of information being accurate, complete, consistent, and non-fictitious, recalled J+ other scholars.

The reliability and credibility of 'FF' reports (PoA) must be noted: PoA website makes them available for 2014-18, with values corresponding in fact to the previous year, as an asterisk (*) notes at page 1 from 2015 to 2018, but not in 2014. However, 'FF 2015' cannot be found in PoA website, where 'FF 2016' is erroneously uploaded, but by Google search. Also, all throughput data for 2003-15 in the table at 'FF 2017' (p. 10) correspond in fact to the period 2004-16, as in its other tables of that document and in all next ones. Values in the original '2011 The port in figures' correspond to the same year 2011. And, all figures are different from a year to the next year, as shown in §4.4.

Other elements related to this cluster of port DQ dimensions will be added to what is said in subsections §4.1-5. It is difficult to know how much the furnished data are unbiased and impartial. However, the fact that SISI has ranked more Chinese ports than other agencies is significant. It is interesting to note that Suzhou and Tangshan which are even not ranked with the major ports by China Statistical Yearbook, but also Rizhao and Yingkou that do since ever, are ranked by SISI, but also by PoR, but not ranked by AAPA. Is there any business relationship between Chinese and Dutch ports? Lacking confirmed information, it is difficult to answer this question.

Instead, more than comments about the reputation of the agencies ranking ports, where there is much to say, this paper's Authors prefer to deliver their own ranking for top 20 European ports, according to data presented by the concerned PAs of European ports (Table 8), using for Piraeus port the data given by PoR. Data for 2014 and 2015 correspond to values appeared in the annual, statistical, or traffic reports of PAs respectively for these years, where available, not to values that may have been changed in the successive reports of the next years.

Table 8. Throughput data and rankings of top 20 European ports and container ports in 2015 (Authors based on PA sources).

	Top 20 European ports				**Top 20 European container ports** (1000 TEUs)			
	2015 (1000 metric tons)		**2014** (1000 metric tons)		**2015**		**2014**	
Rank	Port	TD	Port	TD	Port	TD	Port	TD
1.	Rotterdam	466,363	Rotterdam	444,733	Rotterdam	12,234,535	Rotterdam	12,297,570
2.	Antwerp	208,420	Antwerp	199,012	Antwerp	9,653,511	Hamburg	9,728,666
3.	Hamburg	137,824	Hamburg	145,673	Hamburg	8,821,481	Antwerp	8,977,738
4.	Novorossiysk	127,080	Novorossiysk	121,630	Bremen Ports	5,479,000	Bremen Ports	5,777,000
5.	Amsterdam Ports	96,530	Amsterdam Ports	97,805	Valencia	4,615,196	Algeciras	4,556,465
6.	Algeciras*	91,950	Algeciras*	87,965	Algeciras	4,515,768	Valencia	4,441,949
7.	Ust-Luga	87,868	Marseille	78,520	Felixstowe**	4,042,886	Felixstowe**	4,072,076
8.	Marseille	81,731	Bremen Ports	78,236	Piraeus	3,287,000	Piraeus	3,585,000
9.	Bremen Ports	73,410	Ust-Luga	75,692	Ambarli	3,090,000	Ambarli	3,489,616
10.	Valencia*	69,601	Le Havre	66,905	Marsaxlokk	3,064,005	Gioia Tauro	2,969,802
11.	Le Havre	68,317	Valencia*	66,629	Le Havre	2,559,410	Marsaxlokk	2,869,131
12.	Primorsk	59,606	St. Petersburg	61,178	Gioia Tauro	2,546,805	Le Havre	2,550,199
13.	Grimsby/Immingham	59,103	Grimsby/Immingham	59,370	Genoa	2,242,902	St. Petersburg	2,374,876
14.	Trieste	57,161	Trieste	57,154	Barcelona	1,965,240	Genoa	2,172,944
15.	Constanta	56,337	Constanta	55,642	Southampton**	1,954,060	Zeebrugge	2,046,586
16.	St. Petersburg	51,514	Primorsk	53,656	St. Petersburg	1,715,139	Southampton**	1,895,303
17.	Genova	50,226	Genova	50,968	Zeebrugge	1,568,938	Barcelona	1,893,299
18.	Dunkerque	46,590	Dunkerque	47,100	Mersin	1,470,000	Mersin	1,490,000
19.	Barcelona*	45,921	Barcelona*	45,314	Sines	1,332,200	La Spezia	1,303,017
20.	London	45,430	London	44,489	La Spezia	1,300,442	Sines	1,227,694
Total		1,980,982		1,937,671		77,458,518		79,718,931

* Fishing and supplying traffic are not added to Spanish as to all ports; For UK ports, larger containers than 40' are averagely considered equivalent with 2.25 TEUs.

4.7 PCDQE7: *Consistency*

Consistency specifies that two data values drawn from separate sets do not conflict with one another: it can be curiously simple or dangerously complex, as two data values drawn from separate data sets may be consistent with

each other, yet both can be incorrect, as consistency does not necessarily imply correctness; it is an information free of contradictions or convention breaks; data are consistent if they do not convey content or form heterogeneity (J+). This DQ dimension relates to other ones as *uniqueness*, the extent to which the columns are not repeated, as no entity exists more than once and no duplicate values do to a unique entity; *redundancy*, semantically equivalent data about the same object/event in separate data stores, where concurrent queries produce the same result; referential *integrity*, assigning unique identifiers to objects to simplify data management, but introducing new expectations that an object identifier is used as foreign key within a data set to refer to the core representation that actually exists; *synchronization*, measuring the equivalence of information stored or used in various data stores and systems and making data equivalent; *structured standardization*, presenting all structured attributes in a standard way; *semantic consistency*, using the same values (vocabulary control) and elements to convey the same concepts and meanings in an information object; *structural consistency*, using the same structure, format, and precision for similar attributes or elements, argued J+ referring to other scholars. J+ defined the DQ consistency by *coherence*: the data *comparability* means consistent data allowing comparisons between organizations and over time.

As it may be understood, port throughput data presented in the four Tables 4-7 offer inconsistency in all its dimensions. Rather, port data inconsistency made these Tables exist and this paper see the light.

4.8 PCDQE8: *Usability and Interpretability*

Data usability refers to the extent to which data can be accessed and understood if presented in an intelligible manner; data interpretability refers to the users' ease of understanding and comprehending data that are clear, without ambiguity, in appropriate language, unit, and definitions: correct interpretation is a good presentation providing everything required to the user, including a key or legend for any possibility of ambiguity, as only unambiguous data allow interpretations that are unique, argued J+. Many DQ dimensions relate to the DQ cluster of usability and interpretability: *comparability* (§4.7), the impact of differences in applied statistical concepts and measurement tools/procedures when comparing data between objects or over time for a same object; *presentation - standardization*, allowing formatted data to be presented consistently across different media, with good formats being flexible so to accommodate changes in user needs; - *clarity*, when information is presented so that clearly communicates the truth of data; - *utility*, when information is presented in an intuitive and appropriate way for the task at hand; - *quality*, measuring how information is presented to and collected from users; - *consistency*, when data are always presented in the same format and are compatible with previous data; *appropriateness*, the most important format quality characteristic, depending on two factors of crucial importance, the user and the medium used, which have vastly different abilities to understand data in different formats – it is related to interpretability; *relevance*, when data are applicable and useful for the task at hand and meet current and potential users' needs, being the right kind of information that adds value to the task at hand and is beneficial by providing advantages for its use; applicability; convenience; cohesiveness; complexity; naturalness; flexibility; ubiquity; and portability (J+).

For the usability and interpretability of port throughput data the reader is sent to the above subsections §4.1-7.

4.9 PCDQE9: *Institutional environment*

In this cluster of 6 DQ dimensions that Eurostat (2011) presented as principles of its quality framework (§3), for port DQ purposes related to port rankings it is worth to note the *professional independence* of statistical authorities from other policy, regulatory or administrative bodies, and from private operators, in order to ensure credibility to the port rankings produced by them. Another fundamental principle strongly related to the first is the *commitment to quality* that statistical officers systematically and regularly should demonstrate to identify strengths and weaknesses and continuously improving the process and product quality. As Jugulum (2014) argued, the resolution of data quality issues should encourage various parts of the organization to work together by improving transparency and looking at the ways that link data quality to the process quality through the use of technology.

While the Port of Barcelona's statistics officer argued that port data vary as depending on customer's declaration, there are PAs in Spain as Algeciras and Valencia and elsewhere that do not change their port data year after year. Even in the case of various dependencies, the commitment to quality and professional independence, in combination with statistical processes and a sound methodology (§4.10) are the best way to improve DQ in all its dimensions.

4.10 PCDQE10: *Statistical processes*

Following Jungulum (2014) advice right mentioned, port data quality and port rankings as a particular case are at the heart of statistical processes that Eurostat (2011) considered as its second cluster of DQ principles based on DQ dimensions. It is worth to mention 2 out of 4 DQ principles of this cluster: the *sound methodology*, which underpins quality statistics, requiring adequate tools, procedures, and expertise; and *appropriate statistical procedures*, which underpin quality statistics as implemented from data collection to data validation. "Cooperation with the scientific community is organised to improve methodology, the effectiveness of the methods implemented and to promote better tools when feasible." (Eurostat). We hope this paper will serve to both indicators of a sound methodology and appropriate statistical procedures within statistical processes in port data and rankings.

5. Conclusion and further research

Since 1971, UNCTAD advised PAs on what data should be recorded and how to collect and present them, not suggesting the complete standardization of port statistics, acknowledging that statistical requirements of all ports are not identical, due to the great variety of legal, administrative, and organizational circumstances in which information and statistical systems have to be established and operated. Nevertheless, UNCTAD has suggested uniformity in the methods and forms of collection and presentation of port statistics, using the first data quality dimensions in ports as adequacy and accuracy, availability of accurate and up-to-date information, though not under this label.

Other agencies like BTS suggest (to US ports) the use of standards and policies to ensure and maximize the quality, objectivity, utility, and integrity of port information and these data *criteria* to select throughput indicators: availability, national consistency, timeliness, relevance and clarity, accuracy and transparency. Eurostat developed a statistics code of practice, which together with the general quality management principles form a common quality framework for the ports of EU member countries, among other public agencies. Among 15 principles of this code of practice, one may find five principles based on these DQ clusters of dimensions: relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability, and accessibility and clarity.

This paper aims to field the gap of no research on information quality of port data collected, consumed, produced by data processing, and presented. It selected by literature review on data quality ten DQ (clusters of) elements including DQ dimensions and related problems and applied them on port names and throughput data required for port ranking lists by several international agencies, national institutions and various scholars.

Various shortcomings are observed in the ranking lists selected, prepared by IAPH, Eurostat, SISI, AAPA, PoR, PoA, Lloyd's, ESPO, and scholars, using various sources of information based on data provided by PAs. The data problems strongly relate to 10 DQ dimensions: completeness, availability and accessibility, currency, accuracy, validity, reliability & credibility, consistency, usability & interpretability, institutional environment, and statistical processes, as well as with DQ dimensions related to these like timeliness, up-to-datedness or freshness, punctuality, verifiability, reputation, perception, traceability, comparability, uniqueness, redundancy, integrity, coherence, precision, relevance, security, presentation standardization, clarity, utility, and consistency, professional independence, commitment to quality, a sound methodology, and appropriate statistical procedures.

Among the most frequent problems related to DQ dimensions in port rankings are related to both formats and values of port names and annual throughputs, which do not help the data collection, processing and presentation. Various agencies do not use the same names for the same ports, as they show various annual throughput formats and values for the same ports. Some PAs write data in a way that even Excel does not recognize for further processing. Lists use truncated or rounded values in a various number of decimal places, ones, tens, hundreds, thousands, etc.

Some port ranking lists do not fill the whole port data, as for instance IAPH, Antonellini, etc. ranking 19 out of 20 ports in their lists. For a myriad of unrecognized reasons, some other lists do not include the ports they should, even though these ports, especially from China, India, and Russia, have been since many years recognized by some other lists produced by other agencies. However, China Statistical Yearbook itself did not yet entered ports like Suzhou and Tangshan, while having entered since ever ports like Rizhao and Yingkou, and since 2008 the port of Ningbo-Zhoushan, today's world busiest port. However, Tangshan is part of these cases when a port enter a list and then disappear from it two years after, though having an annual throughput of 109 million tons the preceding year. Some other ports like Nanjing, a river port for CSY, make a one-time entry in a top 100 ports list and directly in

high ranks, 32nd, with 111 million tons, and then disappear again. Other ports as Pyeongtaek, Ust-Luga or Colombo enter in high ranks, respectively in 37th, 46th or 50th place, without having been in lower ranks before, while ports like Yokohama disappear the next year from the 35th place with over 114 million tons, after many years of presence in a top 100 list of world ports, though remaining in the list of top 100 world container ports. All these very curious movements of ports occur in AAPA top 100 world ports and container ports, certainly not being unique cases; other cases are similar for other ports in other ranking lists. The list of port names, figures, and ranking agencies is longer.

Some PAs justify their various port throughput values they give on documents produced in various dates even within a year or more frequently from year to year by dependency on their customer's declarations, while other PAs have final figures since the next year of a given year. BTS urges US ports that their throughput measures should be final and available for the preceding year, while Eurostat let to PAs 15 months to present port throughputs. Etc., etc.

This paper hopes to be followed by others raising issues on how PAs collect, produce and present other port data.

Acknowledgements

We are much indebted to our friends Xhorxhi Qirici and Sulejman Hoxha for their invaluable financial help.

References

AAPA, 2002-2016, World port ranking by total cargo and container traffic. Excel worksheets or .pdf files.

Antonellini, L., 2016, Global Issues: EU port statistics. www.porttechnology.org, edition 71, September, pp. 21-24.

Assoporti, Movements in the principal Italian ports - year 2014, 2015 (in Italian). Elaboration based on data of PAs and ASPO.

Australian Bureau of Statistics, 2009, ABS Data quality framework. May 2009, pp. 16.

Autorità Portuale (AP) di Genova, Traffici ed avviamenti al lavoro del porto: anno 2014-2015. Direzione Pianificazione e Sviluppo, pp. 74, 73.

AP di Gioia Tauro, Medcenter Container Terminal activity - year 2014, 2015. http://www.portodigioiatauro.it/movimenti-medcenter/.

Batini, C., Scannapieco, M., 2006, Data Quality: Concepts, Methodologies and Techniques. Springer, Heidelberg, pp. 275.

Big Port of St. Petersburg, 2012-2017 Грузооборот портов Санкт-Петербург, Приморск, Выборг, Усть-Луга и Высоцк за 12 месяцев 2012-2017 годы (Turnover of St. Petersburg, Primorsk, Vyborg, Ust-Luga, and Vysotsk ports for 12 months 2012-2017), Excel files (in Russian).

Bureau of Transportation Statistics, 2018, Port performance freight statistics - Annual report to Congress 2017, US Department of Transportation.

Clemenson, D., 2017, Tonnage titans - top 20 ports by annual cargo throughput. Fairplay Magazine, October 15, File uploaded on the website of the Port of South Louisiana, pp. 8.

Dunkerque Port, Rapport d'activité 2015 (2016): 2014 (2015) statistics and highlights (bilingual). pp. 69, 64.

ESPO, 2015, Traffic data of year 2014. pp. 431.

ESPO, 2014-2017, Annual report 2014-2015, 2015-2016, 2016-2017. pp. 50-75.

ESPO, 2015, EU ports traffic data. Rapid Exchange System Statistics, https://www.espo.be/fact-and-figures, last access on July 31st, 2018.

Eurostat, 2011, European statistics code of practice. Adopted by the European Statistical System Committee, pp. 8.

Eurostat, 2018, Maritime ports freight and passenger statistics 2016. European Commission, pp. 16.

Eurostat website, Maritime transport (mar) - Eurostat metadata, http://ec.europa.eu/eurostat/cache/metadata/en/mar_esms.htm, last visit on July 31st, 2018.

Garvin, D., 1984, What does product quality really mean? Sloan Management Review (Fall), 25-43.

Garvin, D., 1987, Competing on the eight dimensions of quality. Harvard Business Review (November-December), 101-109.

GPPC, 2015, Newsletter N°12. December, pp. 1.

Greenwald, B., Stiglitz, J., 1990, Asymmetric information and the new theory of the firm: Financial constraints and risk behavior. NBER Working Paper Series, Working Paper No. 3359, pp. 14.

Hansson, S., 1994, Decision theory: A brief introduction. Department of Philosophy and the History of Technology, Royal Institute of Technology, Stockholm, pp. 94.

Herzog, T., 2007, Data quality and record linkage techniques. SOA Annual Meeting & Exhibit, Session 94: Data Quality – Playing with Matches, slides 114.

IAPH, 2015 (2017) World port traffic data for 2005-2014 and 2010-2014 (2007-2016 and 2012-2016). pp. 15 (14).

Jayawardene, V., Sadiq, S., Indulska, M., 2015, An analysis of data quality dimensions. University of Queensland.

Jugulum, R., 2014, Competing with High Quality Data: Concepts, Tools, and Techniques for building a Successful Approach to Data Quality. John Wiley & Sons, Inc., New Jersey, pp. 301.

Lee, Y., Strong, D., Kahn, B., Wang, R., 2002, A methodology for information quality assessment. Information & Management 40.2, 133-146.

Le Havre Port (Haropa), Statistiques définitives 2014, 2015 (mois par mois). Division Statistique, pp. 1, 12.

Lloyd's list, 2017, One hundred [container] ports 2017 [in 2016], Informa UK Ltd., pp. 132.

Mattessich, R., 1993, On the nature of information and knowledge and the interpretation in the economic sciences. Library Trends 41.4, 567-593.

Mersin International Port, 2014-2015 Operational & financial results. pp. 8, 7, https://en.mersinport.com.tr.

Minister of Industry, 2009, Statistics Canada: Quality Guidelines, 5th edition, Catalogue no. 12-539-X, pp. 89.

Moore, G., Benbasat, I., 1991, Development of an instrument to measure the perceptions of adopting an IT innovation. Information Systems Research 2.3, 192-222.

National Bureau of Statistics, 1996, 1998-9, 2000, 2003, 2005-17, Transport(ation) chapter, in: China Statistical Yearbook, pp. 977 (2003).

NCSP (Novorossiysk Commercial Sea Port), Annual financial report 2014, 2015. pp. 64, 135.

Office for National Statistics, 2013, Guidelines for measuring statistical quality, Version 4.1. Sept 2013, UK, pp. 88.

Port de Barcelona, Traffic statistics 2014, 2015. pp. 23.

Port of Ambarli Authority (ALTAS *Ambarli Port Facilities* Trade Company Inc.), website www.altasliman.com.

Port of Amsterdam, 2014, 2015 Jaarverslag (in Dutch). pp. 63, 78.

Port of Antwerp, 2014-2017 Facts & figures. pp. 21 (40 only in 2014).

Port of Antwerp, Statistics yearbook 2014, 2015. pp. 98.

Port of Antwerp, 2011-2016 Annual report. pp. from 23 (2016) to 111 (2012).

Port of Constanta, 2014, 2015 Annual report. pp. 20, 24.

Port of Hamburg Marketing, Annual report 2014, 2015. pp. 56, 60.

Port of Marsaxlokk, http://www.transport.gov.mt/admin/uploads/media-library/files/Cargo%20Throughput%202015.pdf, Cargo throughput at Malta Freeport 2003-2015.

Port de Marseille-Fos, Figures Janvier-Décembre 2014, 2015. pp. 1.

Port of Rotterdam, 2010, 2012, 2014, 2015 Facts & Figures: A wealth of information. pp. 13-22.

Port of Rotterdam, 2016 Port statistics(: A wealth of information). pp. 21.

Port of Sines, 2014, 2015 Traffic statistics. pp. 37, 44.

Port of Trieste, Throughput statistics 2014, 2015. pp. 1.

Port of Valencia, 2014, 2015 Statistical yearbook. pp. 167, 203.

Port of Zeebrugge, 2014, 2015 Jaarverslag (in Dutch). pp. 24, 28.

Puerto de Bahía de Algeciras, 2014, 2015, Resumen tráfico portuario, Periodo solicitado: de January a December de 2014 (2015) comparado con el mismo periodo de 2013 (2014), Excel files (in Spanish).

RAND, 2015, Standards for high-quality and analysis. Report, RAND Corporation, US, pp. 25.

Robinson, R., 2002, Ports as elements in value-driven chain systems: The new paradigm. Maritime Policy and Management 29.3, 241-255.

SISI, 2010-2016 Global port development report. pp. 38 (2010 abstract) and 72-155.

Stigler, G., 1961, The economics of information. The Journal of Political Economy 69.3, 213-225.

Stiglitz, J., 2008, Economic foundations of intellectual property rights. Duke Law Journal 57.6, 1693-1724.

Stiglitz, J., 2016, The revolution of information economics: The past and the future. World Bank, Presentation.

Strong, D., Lee, Y., Wang, R., 1997, Data quality in context. Communications of the ACM 40.5, 103-110.

The Ports of Bremen, Facts & figures Bremen/Bremerhaven 2014, 2015. The Senator for Economic Affairs, Labour and Ports, pp. 32.

UK Department for Transport, 2017, Port freight statistics - Key port statistics 2010-2016. Tables Port0 103, 301, 411, 419, 429, 446, Ods files.

UNCTAD, 1971, Port statistics: Selection, collection and presentation of port information and statistics. Report, TD/B/C.4/79/Rev.1, Genève.

UNCTAD, 1973, Berth throughput: systematic methods of improving general cargo operations. Report, TD/B/ C.4/109 and Add.1, Genève.

Wayback Machine, 2006, Suzhou port prospers, https://web.archive.org/web/20070927062234/http://www.suzhou.gov.cn/news/2006/11/22/eng/eng-9-21-17-3245.shtml.

Wand, Y., Wang, R., 1996, Anchoring data quality dimensions in ontological foundations. Communications of the ACM 39.11, 86-95.

Wang, R., 1998, A product perspective on total data quality management. Communications of the ACM 41.2, 58-65.

Wang, R., Guarascio, L., 1991, Dimensions of data quality: Toward quality data by design. IFRSC Discussion Paper, #CIS-91-06, Composite Information Systems Laboratory, Sloan School of Management, MIT, Cambridge, MA.

Wang, R., Strong, D., 1996, Beyond accuracy: What data quality means to data consumers. Journal Management Information Systems 12.4, 5-33.

Wang, R., Storey, V., Firth, C., 1995, A framework for analysis of data quality research. IEEE Transactions on Knowledge and Data Engineering 7.4, 623-640.

Yokohama Port Corporation, 平成 28 年国内主要港湾統計 (Heisei 28 year [2016] major port statistics), 32 Excel files (in Japanese).

Zhu, H., Madnick, S., Lee, Y., Wang, R., 2014, Data and Information Quality Research: Its Evolution and Future, in: Computing Handbook, 3rd edition, In: Topi, H. (Ed.). CRC Press, Boca Raton, FL, pp. 16.1-16.22.