RESEARCH PAPER

A framework for the quality-based selection and retrieval of open data - a use case from the maritime domain

Milena Stróżyna¹ · Gerd Eiden² · Witold Abramowicz¹ · Dominik Filipiak¹ · Jacek Małyszko¹ · Krzysztof Węcel¹

Received: 21 September 2016 / Accepted: 1 November 2017 / Published online: 16 November 2017
© The Author(s) 2017. This article is an open access publication

Abstract The paper presents a proposal for a framework for the identification, assessment and selection of open data sources based on certain quality criteria, such as accessibility, relevance, accuracy & reliability, clarity, timeliness & punctuality, and coherence & comparability. The framework concerns mainly open data sources and focuses on their quality. The open data are used to enhance existing internal data and to fuse them with data from other sources. The framework consists of few steps starting from definition of quality criteria based on review of relevant literature and user requirements, then identification of potential sources, sources assessment and selection, and finally data retrieval

process. For each step, a specific approach is described, how it may be conducted in practice. The proposed framework is evaluated using a real use case scenario from the maritime domain. The main approach utilized in this use-case is the Delphi method with some characteristics of Analytic Hierarchy Process.

Keywords Internet data sources · Quality assessment · Selection methodology

JEL classification C55 · C88 · D83 · L15 · L86

Responsible Editor: Hans-Dieter Zimmermann

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s12525-017-0277-y>) contains supplementary material, which is available to authorized users.

✉ Milena Stróżyna
milena.strozyzna@ue.poznan.pl

Gerd Eiden
eiden@luxspace.lu

Witold Abramowicz
witold.abramowicz@ue.poznan.pl

Dominik Filipiak
dominik.filiapiak@ue.poznan.pl

Jacek Małyszko
jacek.malyszko@ue.poznan.pl

Krzysztof Węcel
krzysztof.wecel@ue.poznan.pl

¹ Poznan University of Economics and Business (PUEB),
Niepodległości 10, 61-628 Poznań, Poland

² LuxSpace Sarl, SES Business Center, 9, rue Pierre Werner,
L-6832 Betzdorf, Luxembourg

Introduction

Business entities operating in the era of the digital economy need to be supplied with appropriate data in order to make informed decisions and stay competitive. The required data can be acquired from various sources. A source of data can be internal (e.g. transactional data) or external (e.g. sensors, external systems and databases, Internet). Irrespective of the type of data, each data source that is going to be used for decision-making needs to be appropriately defined and assessed in order to ensure the delivery of high-quality results (Robey and Markus 1984). It also has to be relevant for the entity and fulfill certain criteria concerning its purposes and the domain in which it operates.

The goal of this paper is to present a framework for the selection and assessment of open internet sources that can be used as a data source by various business entities. Data from the Internet are used to enhance proprietary data, and are fused with data from other sources, such as legacy systems or internal databases, thus facilitating the entity in conducting its operations based on the broader understanding of its environment.

The general scope of this paper encompasses a procedure for the identification, assessment and selection of internet data

sources. The framework mainly concerns open data sources, and focuses on their quality. The design of the proposed framework was driven by the standard approach to data quality, which defines quality as “the totality of features and characteristics of a product or service that bears its ability to satisfy stated or implied needs” (International Organization for Standardization 1986). In the case of a business entity these needs are expressed in the form of users’ requirements. Therefore, we assume that each potential data source should be analyzed and assessed taking into account two aspects: (1) users’ requirements; (2) a set of selection criteria. Within the framework we define a set of quality measures which are used to assess internet data sources. Then, a process of sources’ assessment, calculation of quality grades and the final selection is specified. Additionally, we discuss what technical requirements are imposed by different types of internet sources when it comes to the data retrieval process, and how such requirements may be fulfilled.

The main research problem tackled in this paper is the development of a quality-based approach for the identification, assessment and selection of internet data sources. As part of a solution we demonstrate how the implementation of such a framework may be conducted in practice. It is non-trivial since it involves handling numerous types of data, very often domain specific. We address this problem by the evaluation of the framework with a use case from the maritime domain. To identify the key factors related to the problem, we defined the following research questions:

- How can we identify a data source relevant to a given problem?
- How can we perform a quality-based assessment of a data source relevant to the problem?
- How can we acquire the relevant data? Which techniques should be used?

We conducted a literature review (see “**Related Work**” section) and observed that the main attention has been paid to concepts of data quality, and the assessment of sources was conducted from this point of view. High-quality sources are usually paid ones, and our motivation was to find equivalent open data sources. Therefore, one aspect of the research was the question: is it possible to use open data instead of paid sources? The combined aspects of access to sources – technical, economic, and legal – were not covered. This constituted the research gap addressed by this paper.

This article is an extended version of a paper presented at the Business Information Systems Conference in 2016 in Leipzig and published by Springer in the LNBIP series (see [Stróżyńska et al. 2016]).

The paper is built around a use case from the maritime domain, briefly described in the next section. Section “**Research methodology**” introduces our research methodology.

Section “**Related work**” presents related work in the area of data quality and data sources assessment. Then, Section “**Framework for the selection of data sources**” describes a proposal for a framework for the identification and selection of internet data sources. In Section “**Framework evaluation using a use case from the maritime domain**” the framework is evaluated using a real use case scenario from the maritime domain, followed by a description of how data from the selected sources can be retrieved. In Section “**Discussion and conclusions**” we summarize the described results.

Use case

System for intelligent maritime monitoring

Nowadays, with growing importance of the maritime trade and maritime economy, one of the key priorities and critical challenges is to improve maritime security and safety by providing appropriate information about the current situation at sea, i.e. by the creation of so-called “Maritime Domain Awareness (MDA)”. This, in turn, can be realized by providing tools which are regularly supplied with high quality data and thus support maritime stakeholders in situation analysis and decision making. The creation of MDA implies the collection, fusion and dissemination of huge amounts of data, coming from many data sources. However, current capabilities to achieve this awareness are still insufficient, resulting in the need to develop dedicated maritime information systems. This need especially concerns systems that are able to fuse in real-time data from various heterogeneous sources and sensors.

This challenge was addressed by the SIMMO project¹. We developed the system aiming at improving maritime security and safety by providing high quality information about merchant vessels (ships that transport cargo for hire) and automatically detecting potential threats. The concept of the SIMMO assumes the constant retrieval and fusion of data from two types of sources:

1. Satellite and terrestrial Automatic Identification System (AIS)², which provides, among other things, information about the location of ships, and generic static information about them.
2. Open internet sources that provide additional information about ships, which is not included in AIS (e.g. flag, vessel type, owner).

The SIMMO system is supplied with information from various internet sources. Therefore, before the system was developed, the appropriate data sources had to be identified

¹ SIMMO – System for Intelligent Maritime MONitoring, the JIP-ICET 2 project financed by the Contributing Members of the European Defence Agency.

² <http://www.imo.org/en/OurWork/Safety/Navigation/Pages/AIS.aspx>

and selected. The initial analysis of the literature in this domain revealed that there is no standard methodology for selecting and assessing the quality of internet sources, particularly for those used in maritime-related systems. The process of developing the SIMMO system is also used as a real use case scenario for evaluation of the proposed framework.

User requirements

The SIMMO project followed a typical approach to modeling and developing an information system, and consisted of the following phases: business analysis, system design, implementation, integration and deployment. The data sources selection was part of the system design and implementation phases. Before that, gathering and analysis of user requirements took place.

The process of user requirements' gathering may be supported by various methods (e.g. interviews or workshops with users, naturalistic observation, competitor analysis, studying documentation) (Rogers et al. 2011). In the SIMMO project firstly we conducted a mail survey..

The mail survey has been conducted using a questionnaire with open-ended questions (see the [Supplementary material](#)). The questionnaire was prepared based on the knowledge obtained from literature review and expertise of team members. The questions aimed at identifying the actual system and data sources used, shortages of existing solutions, potential users as well as defining a process of anomalies detection. When it comes to sample selection, we performed the judgment sampling. The questionnaire has been sent to 5 marine experts, out of which 3 has responded. Based on the provided answers, we were able to elaborate and consolidate various need coming directly from potential users, primarily originating from maritime traffic monitoring activities and anomaly detection practices.

Simultaneously, we have studied documentation regarding existing maritime law, rules and procedures as well as conducted a competitor analysis. This analysis aimed at comparing functionality of systems/applications available on the market. The goal was to evaluate the features, technology, content, usability and overall effectiveness of services available to in the maritime domain. In this process several existing systems and applications were analyzed, what allowed us to identify shortcoming and gaps which could have been addressed by the SIMMO system.

The described process resulted in the specification of requirements for the SIMMO system, which takes into account the challenges and shortcomings of existing surveillance capabilities, extends the current state-of-the-art in maritime surveillance solutions, and addresses the needs of potential users. The requirements were then used in the detailed design of the system. Additionally, these results were used while identifying the potential data sources for the SIMMO system and their assessment according to the proposed framework.

Open data

According to the widely accepted Open Definition “open data and content can be freely used, modified, and shared by anyone for any purpose”³. The concept is not new but it was popularized by open-data government initiatives such as [Data.gov](#) and [Data.gov.uk](#). It was later regulated by the European Commission in the Directive 2003/98/EC on the re-use of public sector information (EU 2003). This directive has an economic goal to facilitate the development of innovative services and free exchange of market information.

Open data is a movement raising interest for its potential to improve the delivery of public services by changing how the governments work. It can also empower citizens and create added value for enterprises. The reports suggest that open data can annually unlock \$3-5 trillion in economic value (Manyika et al. 2013). Further potential can be released by applying advanced analytics to combined proprietary and open knowledge.

Internet sources related to the maritime domain

As mentioned in “[System for Intelligent Maritime Monitoring](#)” section, the SIMMO system should allow for the constant retrieval of data from open internet sources related to ships. Maritime-related data sources can be divided into three categories. The first and most widely used are sensors, which provide kinematic data for the observed objects in their coverage area. The second category includes authorized databases, containing information about vessels, cargo, crew, etc. However, most of the data sources from this category are classified and accessible only to maritime authorities. As a result, they can be referred to as closed data sources. Moreover, most of them are not published in any form on the Internet.

The third category consists of data sources which are publicly available via the Web. Examples of such data sources are portals and ports' websites, where data on vessel traffic is published, as well as blogs, forums and social networks which share information about maritime events (Kazemi et al. 2013). These data include, inter alia, vessel traffic data, reports, press releases and news. In our use case we focus only on data sources from the third category.

In order to identify, assess and select internet data sources for the SIMMO system and then to develop methods for data retrieval, the framework proposed in this paper has been followed.

Research methodology

Since potential use cases for the proposed framework are related to the area of information systems, a methodological review of the available literature was based on an approach

³ <http://opendefinition.org/>

presented by Levy and Ellis (2006). Their methodology is based on three steps of literature review: input, processing, and output. Our study concerns the problem of selecting data sources and assessing them, with an example of applying the proposed framework in the domain of maritime surveillance. Therefore, the general and often interconnected concepts being in the scope of the review are *data quality*, *information retrieval*, and *quality assessment* (for the method itself) along with *Maritime Domain Awareness*, *maritime surveillance*, and *AIS data processing* (for the use case).

We have started the literature review with identification of research related to the use case. For this end, Elsevier, IEEE and SpringerLink databases were used. These sites altogether provided numerous relevant papers to examine. The input stage of conducting the literature review started with a standard keyword search. The initial keywords for the examined use case were “maritime surveillance”, “maritime domain awareness”, “AIS”. Having yielded results, each paper was initially examined. The backward/forward references and author searches were conducted, along with the previously used keywords backward search. The whole process of building the body of knowledge was conducted in a *concertina* manner (constantly narrowing and enlarging the set of papers). Number of returned papers related to our research and their distribution in time is presented in the Fig. 1.

Then, we have carried out a systematic literature review in order to identify the most important aspects concerning the quality of linked data. We have used the Scopus database and searched for the term “quality” in the title only and “linked data” in other textual elements. Out of 68 retrieved articles we filtered out 40 based on title, among which irrelevant medical and health-related articles dominated. The remaining 28 papers can be roughly divided into describing quality features, proposing methods for quality issues detection, and analyzing quality at various levels.

With a good deal of papers collected, we moved to the next stage: processing. The initial work in this step consisted in a rigorous analysis of previous research in order to organize it by similar concepts and findings. This enabled literature synthesis, which is the actual process *en route* to refine the resulting body of knowledge and tailor it for the needs of the presented research. All the presented claims were evaluated, and a selected number of papers was presented in a comprehensive way.

Having done with the literature review both on data quality and the use case, the refined body of knowledge laid the foundations for the presented framework and research. Finally, it was possible to precisely define the research questions and gaps to tackle in the context of the existing work.

Related work

The possibility to use open data in the maritime domain has already been mentioned by Kazemi et al. (2013). They studied the potential to use open data as a complementary resource for anomaly detection in maritime surveillance. As it was an initial idea and realized in the form of a case study, the scope of the research was limited. Three specific open sources were used (including MarineTraffic.com), and the surveillance area was restricted to the region between Sweden, Finland, and Estonia. Moreover, the implemented system was tested by considering one week data. The paper focused on the local perspective, although the potential for a global approach was mentioned. Additionally, some of the methods for data preparation were simplified, for example ships’ names were matched using a string similarity method. In this article we developed a framework for a much larger scale: we collect data from satellite AIS for the whole world since December

Fig. 1 Number of search results related to the use case. Source: own research

2015. Also, we supplement our data with Internet sources as according to the methodology developed by us.

In most cases open data means crowdsourced data, i.e., provided by a community of users. This results in certain disadvantages: quality is mentioned as one of the challenges (Węcel and Lewoniewski 2015). Data may be incomplete, not up-to-date, inaccurate, or incorrect. One of the approaches to mitigate these deficiencies is to use several sources and then verify information.

ISO9000:2015 defines data quality as the degree to which a set of characteristics of data fulfills requirements (Peter 2010). Similarly, the ISO 8402-1986 standard defines quality as “the totality of features and characteristics of a product or service that bears its ability to satisfy stated or implied needs”. Among academics there is no agreed approach to the assessment of data quality.

In the information systems literature, a lot of various data quality attributes can be found. In order to identify the most important aspects concerning the data quality we have carried out a systematic literature review described in the previous section. As a result, we retrieved information about popular quality attributes. We identified a few quality frameworks: SWIQA (Fürber and Hepp 2011), Sieve (Mendes et al. 2012), LiQuate (Ruckhaus et al. 2013), and Luzzu (Debatista et al. 2015).

Batini et al. (2009) presented different definitions of popular quality attributes provided in the literature. Heinrich and Klier (2015) explained the quality of data as a multidimensional construct embracing multiple dimensions, e.g. precision, completeness, timeliness, consistency. Each dimension contributes its own view on the quality of attributes in an information system. In the domain of quality of information a good summary is presented by Eppler, who proposed 70 attributes of information and then narrowed the list to the 16 most important ones (Eppler 2006).

Methods and criteria for quality evaluation also differ in various domains, e.g. business, medical, or technical information. For example, the Commission of the European Communities has elaborated dedicated quality criteria for web pages related to health care. In this case the quality of a web page (effectively, of information) is measured based on the following criteria: transparency; honesty; authority; privacy and data protection; updating of information; accountability; accessibility (Commission of the European Communities 2002).

Taking into account the fact that there is little agreement on the nature, definition, measure and meaning of data quality attributes, the European Parliament decided to propose its own uniform standards for guaranteeing the quality of results for the purposes of public statistics, described in the ESS Quality Assurance Framework (European Statistical System 2014). In this standard, seven quality criteria were defined (European Parliament 2009): (1) *relevance* (the degree to which data meet the current and potential needs of the users); (2) *accuracy* (the closeness of estimates to the unknown true

values); (3) *timeliness* (the period between the availability of the information and the event or phenomenon it describes); (4) *punctuality* (the delay between the date of the release of the data and the target date); (5) *accessibility and clarity* (the conditions and modalities by which users can obtain, use and interpret data); (6) *comparability* (the measurement of the impact of differences in applied measurement tools and procedures where data are compared between geographical areas, sectoral domains, or over time); (7) *coherence* (the adequacy of the data to be reliably combined in different ways and for various uses). In our approach we decided to refer to the above-listed quality attributes.

The quality report according to the ESS should also include additional aspects, such as *Cost and burden* (cost associated with the production of statistical products and burden on respondents), *Confidentiality* (which concerns unauthorized disclosure of data) and *Statistical processing* (operations and steps performed by a statistical system to derive new information). These additional elements were not included in our approach since they seem to be irrelevant in the assessment of open internet sources.

When the quality attributes are defined, the next step is data quality assessment. In this matter the literature also provides a wide range of techniques to assess and improve the quality of data. In general, the assessment consists of several steps (Batini et al. 2009): (1) *data analysis* (examination of data schemas, complete understanding of data and related architectural and management rules); (2) *data quality requirements analysis* (surveying the opinions of users and experts to identify quality issues and set quality targets); (3) *identification of critical areas* (selection of databases and data flows); (4) *process modeling* (a model of the processes producing or updating data); (5) *measurements of quality* (selection of quality attributes and definition of corresponding metrics). The measurement of quality can be based on quantitative metrics, or qualitative evaluations by data experts or users.

In an approach by Dorofeyuk et al. (2004) a data source is described by three qualities in this method: *understandability* (a subjective criterion), *extent* (an objective criterion), and *availability* (an objective criterion), whereas the *efficiency* of a given data source is the weighted sum of its quality scores. Weights are calculated using linear programming. An important feature of this method is the fact that it focuses on each data source selectively. With regard to the step of quality measurement, it can be performed with different approaches, such as questionnaires, statistical analysis and the involvement of subject-matter experts (expert or heuristic techniques).

In the context of open data, it is also useful to look into the quality of linked open data sources. The field of linked data quality is relatively new when compared to well-established publications about data quality, but has an advantage of a structured approach inherent in linked data. Researchers are concerned not only with regards to the quality of the data

sources but also the corresponding metadata, which can compromise the searchability, discoverability, and usability of resources. Neumaier et al. (2016) identified “(meta-)data quality issues in Open Data portals as one of the core problems for wider adoption and also as a barrier for the overall success of Open Data”. The focus of the paper is on open data portals, i.e. portals dedicated to publishing open data using one of the three widely used portal software frameworks (CKAN, Socrata, and OpenDataSoft). The authors leveraged the underlying metadata and mapped it to the standardized Data Catalog Vocabulary (DCAT) metadata schema. Then, they defined a set of quality metrics for this standard. Unfortunately, due to such an approach the reuse is restricted.

Zaveri et al. (2016) carried out a comprehensive systematic review of data quality assessment methodologies applied to linked data. They analyzed in detail 30 papers and 12 tools, identified 18 quality dimensions and 69 metrics. The identified dimensions were clustered into four groups. The following dimensions along with their respective groups are relevant for our approach: availability and licensing (accessibility dimensions); consistency and completeness (intrinsic dimensions); relevancy, trustworthiness, and timeliness (contextual dimensions); interpretability (representational dimensions). Rula and Zaveri (2014) integrated the various data quality dimensions into a comprehensive methodological framework for data quality assessment. They distinguished the following steps: 1. Requirements analysis; 2. Data quality checklist; 3. Statistics and low-level analysis; 4. Aggregated and higher level metrics; 5. Comparison; 6. Interpretation.

Framework for the selection of data sources

“Data on the Web reveals a large variation in data quality” (Zaveri et al. 2016). This section presents a framework for the identification, assessment and selection of internet data sources to be used along with data from existing sources (e.g. internal data) in the decision-making process. The proposed framework consists of the following steps: (1) identification of potential data sources; (2) the assessment of data sources, including the definition and selection of quality criteria, and the final selection of sources for a system; (3) the design and development of the data retrieval procedure, including the definition of the cooperation model, the development of data acquisition methods and data fusion. The steps of the framework are presented in Fig. 2 (numbers refer to subsections).

Identification

The first step focuses on the identification of potential data sources available on the Web and related to a given domain or a given issue.

The Web can be divided into two types: shallow and deep. The former is that portion of the Web that is indexable by conventional search engines and links billions of HTML pages. The latter consists of on-line databases that are accessible via Web interface to humans but poorly indexed by regular search engines and, in consequence, not available through a regular web search (Kaczmarek and Wękowski 2013). It is estimated that the deep Web contains a far more significant amount of information, which is “hidden” behind the query forms of searchable databases. Such Web pages are not directly accessible through static URL links but instead are dynamically generated as response to queries submitted through the query interface of an underlying database (Chang et al. 2004). Consequently, both the shallow and the deep Web should be considered as potential data sources.

For identification various search engines can be used, including traditional ones like Google, Bing, as well as meta-search engines or domain-specific ones (if they exist). Apart from the search engines, a review of relevant literature should also be conducted, since there might be information or directions as to what data sources are used in a given domain. Finally, (if possible) consultation with domain experts or future users should also be conducted, who may also suggest potential sources of data (He et al. 2007).

From the data availability perspective, internet data sources can be divided into three categories:

- Open data sources, in which data is freely accessible and reusable to the public (no authorization required).
- Open data sources with required authorization and free registration.
- Closed data sources with required authorization and non-free access.

Quality measures

The identified data sources should be assessed from the point of view of the quality of a source as well as its compatibility with the users’ requirements. At first, the assessment procedure assumes the definition and selection of quality criteria. This selection may result from uniform definitions or standards which are used in a given domain, if such a standard exists.

In our approach we decided to adopt the data quality measures proposed by the European Statistical System (ESS) (2014). This selection was driven by the fact that in the case of maritime systems, there are no standards or procedures which would suggest or dictate the set of quality criteria to be used. Previous research on data quality in the maritime domain mainly concerned the quality of AIS data (such as completeness of AIS messages, their accuracy, integrity, etc.) (MMO 2013; Harati-Mokhtari et al. 2007; Iphar et al. 2015). However, the methodologies and quality attributes

Fig. 2 Source selection method

proposed there do not fit well with the assessment of other maritime-related sources of data, especially data published on the Internet. As a result, we decided to look for commonly used approaches in other domains, like those used in statistics.

Data quality is of primary importance in statistics. Therefore, it is common to follow quality assurance frameworks. In the case of the ESS the unified quality assurance framework proposes well-defined quality attributes. We are aware that this framework is used mainly by the statistical systems and originally may not fit well with quality assurance in all domains. Nevertheless, we believe that these basic quality criteria are universal, and after minor modifications may be used in various research. In our case, we made some modifications to adopt them to the characteristics of the internet data sources.

In the practice of Eurostat, some of the quality criteria were combined (e.g. accuracy & reliability, timeliness & punctuality) (European Statistical System 2014). In our research we did the same, which finally resulted in six quality measures. In the next paragraph we provide a short definition of each measure that is used in the framework:

- *Accessibility* – the possibility to retrieve data from a source; it includes such aspects as the structure of a source, the technologies used, the form of data as well as source stability (changes of a structure, errors, unavailability of a service); it also takes into account terms of use, privacy policy, requirements for login or registration, access to data (fees, subscriptions), etc.
- *Relevance* – what kind of information is provided by a source and whether this information matches the users' or system's requirements.
- *Accuracy & Reliability* – reliability of information provided from the point of view of the users' requirements; it also evaluates data scope and coverage (how much information is available) and data accuracy (missing information).
- *Clarity* – availability of an appropriate description or explanation of the data model and information about a source of information (data provider).
- *Timeliness & Punctuality* – data update (time interval between the occurrence of an event and availability of the data which describe it) and time delay in publishing updated information.

Fig. 3 Pipeline of data acquisition from AJAX and deep Web data sources

- *Coherence & Comparability* – whether data provided in a source describe the same phenomenon or have the same unit of measure like data from other sources.
- The Delphi study is flexible in its design and allows, for example, the involvement of experts from different domains and having different research experience.

Assessment and selection

When the quality measures are defined, the next step is assessment of the potential internet data sources according to these measures. A systematic approach should be followed. In the framework we propose to use expert knowledge (domain experts assess and select data sources) by conducting the assessment based on the Delphi method with elements of the Analytical Hierarchy Process (AHP) proposed by Saaty (1990).

The Delphi method (Brown 1968) relies on a group of experts; its aim is to achieve the most reliable consensus on a given issue. The method is a systematic approach – it consists of rounds in which the experts answer questionnaires and provide their judgements on a given topic. After each round, a facilitator provides an anonymous summary of the experts' opinions. In the next round the experts are encouraged to revise their earlier answers in the light of the replies of other experts. The process is continued until a consensus or a predefined stop criterion is reached (e.g. number of rounds).

The main characteristics of the method are (Linstone and Turoff 2002): feedback of individual contributions of information and knowledge; assessment of the group judgement; opportunity for individuals to revise their statements, and anonymity for individual responses to prevent the authority of some experts from dominating others. Researchers have applied Delphi to a wide variety of situations as a tool for experts solving problem. In IS research Delphi was used mainly for forecasting purposes or in concept/framework development (Okoli and Pawlowski 2004).

We selected Delphi for the following reasons:

- It has proven to be a popular tool in information systems research, employed primarily in cases where judgmental information is indispensable.
- A group study (like Delphi) more appropriately answers the research questions than any individual expert's responses. Among other group decision analysis methods (such as the nominal group technique and social judgement analysis) Delphi is desirable since it does not require the experts to meet physically, which could be impractical for international experts.
- The requirements on the number of experts involved is not precisely defined and rather modest.
- It allows serving a dual purpose: to solicit opinions from experts about potential sources and having rank them according to their importance or appropriateness for further studies.

There are variations of Delphi tailored to specific problem types and outcome goals. In our approach we enhanced the standard Delphi with some characteristics of AHP, which allows the adding of priorities (weights) to the decision-making factors (in our case quality measures). Moreover, following AHP, the experts in Delphi are asked to evaluate a data source under a particular criterion using a four-level rating scale (high, medium, low and N/A), which is then converted into numerical values (accordingly high = 3, medium = 2, low = 1, N/A = 0) and normalized. Based on these evaluations, a final quality grade for each source is calculated.

The rate N/A (not available) means that information required for a particular criterion (e.g. update interval or data coverage) is not specified by a source and, as a result, it was not possible to assess a source in this matter. In the case of the *Accessibility* measure, the rate N/A means that due to the terms of use or privacy policy, it is prohibited to automatically retrieve or use data published in a given source⁴.

Having calculated the quality grade, the final selection of data sources for a given system can take place. Here, we propose to define a threshold value for the quality grade above which the source is selected. In order to define the threshold, again the Delphi method may be used.

Retrieval

In the final step of the framework, the design and development of the data retrieval procedure is foreseen. It includes the definition of the cooperation model, the development of data acquisition methods, and data fusion.

The cooperation model should present how cooperation with a data provider (the selected data source) will look, including: selection policy (which information will be retrieved), re-visit policy (time intervals between data updates), and politeness policy (suggestion for the interval of visits to the same server to avoid source overload). To this end, each source has to be analyzed with regard to the existence of the politeness policy or the terms of use.

Then, an appropriate method for the automatic acquisition of data from the source needs to be implemented. The method depends mainly on the type of source and format of publishing data. In “*Data Fusion*” section examples of retrieval methods for four different categories of data sources are presented.

⁴ In such sources the following provisions are included: “no part of the information contained in the website may be stored in a retrieval system” or “the use of a web-robot or similar techniques to download data in an automated or regular manner is strictly prohibited”

In the case when data are obtained from many heterogeneous sources, they have to be fused, i.e. a common data model meeting the initial system requirements has to be developed and used to organize new data in a homogeneous and integrated form. This entails semantic interoperability problems related to the interpretation of data coming from different sources.

Framework evaluation using a use case from the maritime domain

In this section, the application of the framework presented in the previous section for the SIMMO use case is presented.

Identification of internet data sources

As indicated in “[Framework for the Selection of Data Sources](#)” section, the first step of the framework is the identification of potential sources. In the SIMMO case, potential data sources related to maritime surveillance were identified using search engines, literature reviews and consultations with subject matter experts. The search engines encompassed conventional search engines (like Google) as well as meta search engines like Dogpile⁵, Mamma⁶, and Webcrawler⁷. Apart from the search engines, other data sources were also analyzed, including sources indicated in (Kazemi et al. 2013), and those suggested by maritime practitioners. The other methods were used mainly to identify potential deep Web sources.

As a result, 59 different data sources available on the Web were found. The identified sources were part of both the shallow (22%) and the deep Web (78%). They provided information in a structured, semi-structured and unstructured manner. The list of identified internet data sources is presented in Table 2. From the point of view of data access, we divided them into four categories:

1. Open data sources (O) – websites that are freely available to internet users.
2. Open data sources with registration (OR) – websites that provide information only to authorized users.
3. Data sources with partially paid access (PPA) – websites that after the payment of a fee provide a wider scope of information.
4. Commercial (paid) data sources (PA) – websites with only paid access to the data (fee or subscription required).

⁵ <http://www.dogpile.com/>

⁶ <https://mamma.com/>

⁷ <http://www.webcrawler.com/>

From all the identified sources, for further analysis we selected only open data sources (category O and OR). At this stage, we eliminated commercial data sources and websites with paid access (categories PPA and PA). The elimination of these sources resulted from the fact that they provide only very general, marketing information about the data they have, and access to the data is available only after paying a fee or signing a contract. Moreover, our attempt to make contact with these data providers in order to get access to sample data failed (requests for data access were sent but with no response). Furthermore, in the project we did not foresee buying access to maritime data. Eventually, only sources with public content were selected for the project. Nevertheless, we believe it is sufficient to meet the users’ requirements, and provides the advantages of open data presented in “[Open Data](#)” section.

Similarly, two other data sources (IALA, SafeSeaNet) were rejected due to the fact that access to the data required the application of a long-lasting procedure with no guarantee that access would be granted. Due to the project’s limited duration, there was not enough time to apply for the data. However, in the case of getting access these sources can still be assessed according to the framework and included in the system in the future.

As a result of initial selection, 43 sources were taken into account as potential sources for the SIMMO system and assessed by the experts.

Assessment of internet data sources

In order to select sources of the highest quality and best suited to the users’ requirements, the identified data sources were assessed using the six quality criteria presented in the previous section. Definitions of these criteria were adjusted to the specifications of the SIMMO project (see Table 1).

The process of data source assessment was conducted using the Delphi method. In fact Delphi was utilized three times: for weights assignment, source assessment, and threshold specification. In all cases the same group of experts was involved, consisting of 6 people. The experts were drawn from both inside and outside the project. They were experts either in the maritime domain or in the design and development of information systems (including maritime systems), having experience in data retrieval from various data sources (including structured- and unstructured internet sources). In the selection of experts we followed guidelines on how to conduct a rigorous Delphi study provided by (Hsu and Sandford 2007; Kobus and Westner 2016; Schmidt 1997).

At first, the Delphi method was used to define the importance of the selected quality attributes (by assigning them weights) and thus prioritize the selection criteria. Here, a variant of Delphi called “ranking-type Delphi” was used, which allows the development of a group consensus about the relative importance of issues (Kobus and Westner 2016; Schmidt

Table 1 Quality measures used to assess Internet data sources

Name	Description	Weight
Accessibility	A possibility to retrieve data from a source; website structure and stability	0.3
Relevance	How well the data are fitted to the SIMMO system's requirements	0.3
Accuracy & Reliability	Data scope, Missing elements, Ship coverage	0.2
Clarity	Explanation of source's metadata model, Data provider	0.1
Timeliness & Punctuality	Data update, Time delay in publishing the data	0.05
Coherence & Comparability	Definition of a described phenomenon and units of measure	0.05

Source: authors' own work

1997). This process consisted of three rounds, after which the consensus was reached.

Then, Delphi was used in the process of the assessment of the identified data sources according to the defined quality criteria. In this case the process consisted of two rounds. At the beginning each expert received the gathered basic information about a source and some statistics. Based on this information, as well as their knowledge and experience, they were asked to initially assess each potential data source by assigning a mark to each quality criterion using a four-level rating scale: high, medium, low, N/A, and provide a short justification. Here a questionnaire with a list of sources were used (similar to that presented in Table 2). Then, the results were summarized by a facilitator, and in the second round the experts were asked to review the summary of results and revise their assessments. After this round a consensus began to form, and based on the revised judgments the final marks in each criterion were selected (by majority rule). The results of the quality assessment for each source are presented in Table 2.

Final selection of sources

After the assessment, the final selection of sources took place. Firstly, all sources with *Accessibility* measure marked as *N/A* were removed (12 sources from the O and OR categories, see Table 2). This elimination resulted from the reasons indicated before, regarding access to the data and the prohibition of using information from these sources indicated by the data provider. Also, the sources with *Accessibility* assessed as *Low* were eliminated (5 sources, see Table 2). This encompasses the sources with unstructured information (e.g. text in a natural language). We excluded them due to the fact that while defining the requirements for the system it was decided to include only sources with structured or semi-structured information. The reason for this was the limited time frame of the project and fact that an automatic retrieval of unstructured information would require a significant amount of work on developing methods for Natural Language Processing.

The sources with *Relevance* measure graded as *Low* were also eliminated (12 sources, see Table 2). It was pointless to

retrieve data that are not well-suited to the requirements defined for the SIMMO system. For example, the SIMMO system focuses only on collecting and analyzing data about merchant vessels, and therefore some categories of sources may have been excluded (e.g. fishing vessels, oil platforms).

In the next step, each quality measure was converted into a numerical value: High = grade 3, Medium = grade 2, Low = grade 1, N/A = grade 0. Then, a final quality grade was calculated according to the formula:

$$X_s = \sum_{i=1}^n \frac{x_i}{3} w_i * 100\%,$$

where s means the number of the analyzed sources, $n=6$, x_i means the grade assigned by the experts to a given quality measure i , and w_i means the measure's weight. The grade was also normalized to the range 0-100% (therefore each assigned measure is divided by 3).

Based on the calculated quality grades, a ranking of sources was created. Then, the experts were asked to decide on the threshold for the final selection of sources. After two rounds of Delphi the threshold was set at 85%. From the ranking list only sources with a final grade above the defined threshold were selected for use in the SIMMO system (the one in bold in Table 2).

To sum up, the application of the proposed framework for data source selection in the SIMMO use case allowed us to identify, assess and finally choose open internet data sources of the highest quality, which were then used by the SIMMO system.

Model of cooperation with data owners

In the next step, a model of cooperation with external data providers was defined. By external data providers we understand the sources selected for the SIMMO system. For each selected source a separate cooperation model was designed and described in the documentation. In defining the model, the following aspects were taken into account:

- Scope of available information – what kind of information is available in a source.
- Scope of retrieved information – which information pieces will be retrieved from the source.

Table 2 List of assessed Internet data sources

Type of information	Source name	Type	A	C	R	TP	CC	AR	Quality grade	Selected
General vessel data	Marine Traffic	PPA	H	M	H	H	H	H	98%	Yes
	US Coast Guard	O	H	H	H	H	M	H	98%	Yes
	Maritime mobile Access and Retrieval System (ITU MARS)	O	H	H	H	H	M	M	92%	Yes
	Maritime-connector	O	H	H	H	N/A	H	H	90%	Yes
	ShipFinder	O	H	M	M	H	M	L	73%	No
	AIS HUB	O	H	M	M	M	M	L	70%	No
	Equasis	OR	N/A	H	H	H	M	H	68%	No
	IMO GISIS	OR	L	H	H	N/A	M	H	68%	No
	Vessel finder	O	N/A	L	H	H	H	H	67%	No
	FleetMon	OR	N/A	M	H	H	M	H	67%	No
	Lloyd's Register Ship in Class	OR	M	H	M	L	M	M	65%	No
	ShipSpotting	OR	N/A	H	H	H	H	L	57%	No
	World Shipping Register	PPA	N/A	M	M	M	H	H	55%	No
	IHS	PA	-	-	-	-	-	-	-	No
	Clarkson	PA	-	-	-	-	-	-	-	No
	Internet Ships Register	PA	-	-	-	-	-	-	-	No
	Grosstonage	PA	-	-	-	-	-	-	-	No
	Lloyd's List Intelligence	PA	-	-	-	-	-	-	-	No
	Vessel Tracker	PA	-	-	-	-	-	-	-	No
	International Association of Lighthouse Authorities (IALA)	PPA	-	-	-	-	-	-	-	No
	SafeSeaNet Vessel Traffic Monitoring and Information System	OR	-	-	-	-	-	-	-	No
Ship owners	InfoMare	O	H	L	M	N/A	M	N/A	55%	No
	Seaagent	PPA	N/A	L	H	N/A	L	N/A	33%	No
Weather	ICM Meteo	O	L	H	L	M	L	L	40%	No
	Meteooffice	O	L	H	L	M	L	L	40%	No
Classification of ships	Sailwx	O	N/A	L	L	M	L	M	33%	No
	International Association of Classification Societies (IACS)	O	H	H	H	H	H	H	100%	Yes
	Vessel in class									
	American Bureau of Shipping (ABS)	O	H	H	H	M	M	M	88%	Yes
	International Association of Classification Societies (IACS)	O	H	H	M	H	M	M	82%	No
	Transfer of Class									
	ClassNK	O	N/A	H	H	M	M	M	58%	No
	Leonardo Info	OR	N/A	H	H	N/A	M	H	58%	No
	Bureau Veritas Group	PA	-	-	-	-	-	-	-	No
	China Classification Societies	PA	-	-	-	-	-	-	-	No
	International Register of Shipping	PA	-	-	-	-	-	-	-	No
PSC / Banning / Detentions	Thetis Company Performance	O	H	H	H	H	L	H	97%	Yes
	Tokyo MoU	O	H	H	H	H	L	H	97%	Yes
	Mediterranean MoU	O	H	H	H	H	L	H	97%	Yes
	Black Sea MoU	O	H	H	H	H	L	H	97%	Yes
	Government of Canada - Port State Control	O	H	H	M	H	H	H	90%	Yes
	Indian Ocean MoU	O	M	H	H	H	L	H	87%	Yes
	Riyadh MoU	O	M	H	H	H	L	M	80%	No
	Latin America MoU	O	M	M	H	H	L	M	78%	No
	Paris MoU	O	N/A	H	H	H	L	H	67%	No
	Abuja MoU	O	N/A	H	H	M	L	H	63%	No
Maritime crimes	Caribbean MoU	O	N/A	H	H	N/A	L	H	57%	No
	ICC Commercial Crime Services	PPA	M	H	L	H	L	M	60%	No
Tankers	Maritime Safety Information	O	L	H	L	M	L	H	53%	No
	Q88.com	PPA	H	H	H	M	M	M	88%	Yes
	Auke Visser's International Supertankers	O	L	M	H	L	M	M	63%	No
	International Association of Independent Tanker Owners	PA	-	-	-	-	-	-	-	No
Container ships	Containership-info	O	H	M	L	M	M	H	73%	No
Fishing vessels	Commission for the Conservation of Antarctic Marine Leaving Resources	O	H	H	L	H	H	M	73%	No
	International Convention for the Conservation of Atlantic Tunas (ICCAT)	O	H	H	L	N/A	H	H	70%	No
	Indian Ocean Tuna Commission (IOTC)	O	H	H	L	N/A	H	H	70%	No
	Western & Central Pacific Fisheries Commission (WCPFC)	O	H	H	L	N/A	H	M	63%	No
	FAO Vessel Record Management Framework (FVRMF)	O	M	H	L	L	M	H	62%	No
LNG vessels	Zeus Intelligent	PA	-	-	-	-	-	-	-	No
	LNG World	PA	-	-	-	-	-	-	-	No
Oil platforms	Oil and gas: offshore maps in UK	O	H	H	L	H	H	H	80%	No

Legend: *A* Accessibility, *C* Clarity, *R* Relevance, *TP* Timeliness & Punctuality, *CC* Coherence & Comparability, *AR* Accuracy & Reliability, *H* High, *M* Medium, *L* Low, *N/A* Not available, *O* Open, *OR* Open with registration, *PPA* Partially paid access, *PA* Paid access

- Type of source – whether retrieved content is published in the shallow or deep Web, and in what form data are available, e.g. internal database, separate xls, pdf or csv files.
- Update frequency – how often information in a source is updated; whether the whole content is updated or only new information appears.
- Politeness policy – what kind of robot exclusion protocol was defined by the website administrators, for example which parts of Web servers cannot be accessed by crawlers, as well as requirements on time delay between consecutive requests sent to the server.
- Re-visit approach – how often the SIMMO system will retrieve information from a given source, i.e. the intervals between consecutive downloads from the source, taking into account the politeness policy, if defined.

Retrieval of data from internet sources

Finally, data from the selected sources was to be retrieved, merged and stored in the system for further analysis. The data were to be acquired automatically by the developed Data Acquisition Modules (DAMs). DAMs connect to the data source in a defined manner, send appropriate requests, collect the returned documents and extract the required data.

Each source may have a different structure and may publish data in different ways. If DAM is to successfully acquire the data from a given source, a specific set of technical requirements must be met. Four general categories of data sources were identified in terms of such requirements: (a) shallow Web sources, (b) deep Web sources, (c) sources publishing data in XLS/CSV files, (d) sources publishing data in PDF files.

Below we describe these categories in detail and discuss how data is retrieved in the SIMMO system.

Shallow web sources publish their data in the form of web pages (HTML documents), which can be directly fetched using GET queries defined according to HTTP⁸. As a result, the source sends back an HTML document with data embedded in it. Such documents usually contain data concerning a single entity (e.g. a single ship) or a list of links to web pages that contain data on single entities. The data itself may be extracted from the document using regular or XPath expressions⁹. In order to conduct monitoring of new or updated data published in the source, it is crucial to maintain a list of known URLs of documents published in this source and to manage a queue in which these URLs are to be visited.

For each shallow Web source used in SIMMO, a separate DAM was prepared that was responsible for the actual retrieval and processing of data from a given source. These modules share some common operations, such as queuing mechanisms, retrieval of HTML documents under a given URL, and writing the data to the database. Still, operations such as extraction of the data from the HTML document have to be implemented separately for each source, which is the consequence of different structures of the HTML documents.

Deep Web and AJAX data sources also publish their data in the form of HTML documents, but these documents are not directly accessible through static URL links. Instead, they are dynamically generated in response to queries submitted through the query interface to an underlying database. In order to fetch the data published in sources belonging to this category, DAMs need to perform many additional operations compared to shallow Web sources, such as posting filled forms or executing JavaScript code embedded in HTML documents.

This functionality was implemented with the Selenium WebDriver¹⁰ toolkit and the Mozilla Firefox web browser. The toolkit allows the automation of actions within web browsers. It is then possible to automatically submit instructions to one of the supported web browsers. In our case, the developed DAM (written using the Python language) opens a Mozilla Firefox browser window inside the X virtual framebuffer (XVFB)¹¹. The process of acquisition of the data using the pipeline is presented in Fig. 3.

Data sources with CSV and XLS files form a third category of data sources. CSV (Comma Separated Values) files are regular text files used for storage of tabular data, where each line contains a single record, and fields in the record are separated using a selected separator (usually a comma, hence the name of the format) and quoted. CSV file format can be easily processed by any programming language. Another format very similar to CSV file, in terms of the processing pipeline, is the XLS(X) file type. This is a file format for the representation of spreadsheets.

In the data sources used in the SIMMO use case, CSV and XLS(X) files with the required data are published on a regular basis, e.g. once a week, under a certain URL.

Sometimes, these files are additionally archived, e.g. into a zip file. To fetch the data from these sources, Python scripts were developed which are executed regularly by Cron¹² and monitor a given source to identify if a URL to a previously unseen CSV/XLS file appears on a web page. Once the file is downloaded (and unpacked if necessary), it can be programmatically read and its content can be processed sequentially, row by row, to get data about specific entities.

¹⁰ <http://www.seleniumhq.org/>

¹¹ The XVFB is software which allows the running of applications with the Graphical User Interface on computers without display hardware and physical input devices; see <http://www.x.org/archive/X11R7.6/doc/man/man1/Xvfb.1.xhtml>

¹² Cron is a job scheduling daemon for the Linux operating system

⁸ <https://www.w3.org/Protocols/rfc2616/rfc2616.html>

⁹ <http://www.rfc-base.org/txt/rfc-5261.txt>

Fig. 4 Retrieval of data from sources that publish data in the form of PDF files

Data sources with PDF files are web portals in which data can be accessed by downloading and displaying PDF (Portable Document Format) files accessible under a certain URL. While PDF has some advantages¹³, this format is very difficult for automatic processing. This is due to the fact that it was created to be read by humans. Processing PDF documents becomes even more difficult when we aim at extracting data from a table that is embedded in it in an automatic manner.

The processing pipeline for fetching and processing PDF files is presented in Fig. 4. First, a PDF file is downloaded from the source to the local disk. Next, the file is converted to XML using the `pdftohtml`¹⁴ program, included in the Ubuntu Linux operating system. This program, when executed with the `-xml` option, produces an XML document containing a text which is suitable for further processing. In the obtained XML document, each piece of text is contained in a separate element with data about its coordinates on the document page (i.e. number of units in relation to its top-left corner). A set of manually-crafted rules needs to be developed in order to, based on those coordinates, recreate the original structure of the table and extract the data.

Data fusion

Data obtained using DAMs in the way described in the previous paragraphs are to be stored in an internal database and used in further analysis. Still, there are some challenges that must be dealt with before such data may be used. This results from the fact that data in different sources are not consistent with each other, for example:

- The same entity (e.g. a vessel) in two different sources may be referred to using different names, e.g. different spellings of the name of the vessel, or using different attributes.

- The same attribute for a given entity may have different, conflicting sources in different sources.
- The same attributes may be described using different units of measure (e.g. meters vs feet).

Such situations should be automatically resolved if the system is to be able to utilize the data retrieved from different sources. This process is called data fusion. In case of the SIMMO system, this problem was mainly resolved firstly by assignment of artificial, unique identifiers to each entity and then development of methods that automatically assign these identifiers to each data item related to a given entity. In the proposed methods various approaches are used inter alia text similarity measures, heuristic methods, prioritizing of data sources, analysis of agreement between different attributes, lexicon building based on information provided by DBpedia. Still, this issue is well beyond the scope of this paper. The detailed results of our work on data fusion are described in another paper [Małyżsko et al. 2016].

Discussion and conclusions

The goal of this paper was to propose a framework for the selection of open data sources which provide data to be fused with internal data and with data coming from other types of sources (e.g. sensors). The framework concerns internet sources and focuses mainly on the quality of data sources.

In our paper we refer to the value of data in the context of networked business. We asked if there are open data sources that can be used instead of costly, closed and proprietary sources. In order to use such open sources, we need to evaluate the usability of such sources. Therefore, specific characteristics have to be identified, e.g. quality and availability. The main motivation for the research is how to use open data sources to increase the value of existing closed or internal data sources. Thus, we contribute to the electronic market in a broader sense where access to appropriate data sources impacts the value chain of information usage. We help to answer

¹³ Files in this format are highly portable and easy to open for displaying in different applications

¹⁴ <http://linux.die.net/man/1/pdftohtml>

the important question of what kind of information is available on the market concerning specific domains by providing the framework.

In a nutshell, the proposed framework consists of 3 steps. A dedicated method is developed for each step. Firstly, the framework gives directions on how potential open data sources related to a given problem may be identified. Then, it proposes a method for sources' assessment taking into account a defined set of quality measures. Eventually, it describes how to perform data retrieval and what techniques may be used for this end.

Although the steps of the framework may seem to be standard and already used in many organizations, we believe our approach is innovative and different from the state-of-practice methodologies for several reasons. First of all, as indicated in "Related Work" section, there are a number of quality attributes that are used in practice, but there is no standard set of measures which may be used in all domains. Instead, most often the appropriate attributes are selected taking into account some domain-specific aspects and requirements. In the case of the proposed framework, we faced the problem of the quality assessment of internet data sources, which then need to be fused with other types of data (internal and sensor data). However, instead of selecting our own set of quality attributes, we decided to adopt the measures that are commonly used in another domain (the standard of the European Statistical System) but so far have not been used for the analysis of internet data sources. Another advantage of our approach is that it allows the performance of quality-based assessment of all potential data sources from the domain. At the same time, the assessment procedure may be reused for new potential data sources which appear in the future, using the same set of quality measures and quality threshold. Thanks to this we can make sure that even in the case of the inclusion of new sources, the appropriate quality will still be ensured.

The framework was evaluated using the real use case scenario from the maritime domain. Nevertheless, we also see its potential application in other domains which require the acquisition of data freely available on the Internet. Thus, we encourage other researchers to use the proposed framework to perform quality-based assessments of internet sources related to other domains.

The conducted evaluation gave us an overview on the scope of data related to vessels that is available on the Web and can be freely used in the maritime domain. On the other hand, the conducted analysis revealed that there is a lot of data sources with valuable information that unfortunately cannot be used due to strict terms of use or policies which prohibit using any technique for automatic retrieval of data published in a given source (ca. 28% of all the assessed sources). There are also sources that require prior written authorization to use their data. Such restrictions influence the availability and thus the quality of a source. Ultimately, such sources cannot be perceived as fully open.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3), 16:1–16:52. <https://doi.org/10.1145/1541880.1541883>.
- Brown, B.B. (1968). *Delphi process: A methodology used for the elicitation of opinions of experts* (No. RAND-P-3925). Santa Monica: RAND Corp..
- Chang, K. C.-C., He, B., Li, C., Patel, M., & Zhang, Z. (2004). Structured databases on the web: observations and implications. *ACM SIGMOD Record*, 33(3), 61–70.
- Commission of the European Communities. (2002). eEurope 2002: quality criteria for health related websites. *Journal of Medical Internet Research*. <https://doi.org/10.2196/jmir.4.3.e15>.
- Debattista, J., Lange, C., & Auer, S. (2015). Luzzu - A framework for linked data quality assessment. In ISWC 2015 Posters and Demonstrations Track (Vol. 1486). Bethlehem, United States: CEUR-WS. Retrieved from <http://www.scopus.com/inward/record.url?eid=2-s2.0-84955600970&partnerID=tZOTx3y1>.
- Dorofeyuk, A. A., Pokrovskaya, I. V., & Chernyavskii, A. L. (2004). Expert methods to analyze and perfect management systems. *Automation and Remote Control*, 65(10), 1675–1688. <https://doi.org/10.1023/B:AURC.0000044276.99273.a0>.
- Eppler, M. J. (2006). Managing information quality: Increasing the value of information in knowledge-intensive products and processes (2nd ed.). Berlin Heidelberg: Springer-Verlag. doi:<https://doi.org/10.1007/3-540-32225-6>.
- EU. (2003). Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information. *Official Journal of the European Union*, 46(L 345), 90–96.
- European Parliament (2009). Regulation (EC) No 223/2009 of the European Parliament and the Council of 11 March 2009 on European statistics and repealing Regulation (EC, Euratom). *Official Journal of the European Union*, 52.
- European Statistical System (2014). ESS handbook for quality reports. Eurostat.
- Fürber, C., & Hepp, M. (2011). SWIQA - A Semantic Web information quality assessment framework. In 19th European Conference on Information Systems, ECIS 2011. Retrieved from <http://www.scopus.com/inward/record.url?eid=2-s2.0-84870643288&partnerID=tZOTx3y1>.
- Harati-Mokhtari, A., Wall, A., Brooks, P., & Wang, J. (2007). Automatic Identification System (AIS): data reliability and human error implications. *Journal of Navigation*, 60(3), 373–389.
- He, B., Patel, M., Zhang, Z., & Chang, K. C.-C. (2007). Accessing the deep web. *Communications of the ACM*, 50(5), 94–101.
- Heinrich, B., & Klier, M. (2015). Metric-based data quality assessment — Developing and evaluating a probability-based currency metric. *Decision Support Systems*, 72, 82–96. <https://doi.org/10.1016/j.dss.2015.02.009>.
- Hsu, C.-C., & Sandford, B. A. (2007). The Delphi technique: making sense of consensus. *Practical Assessment, Research & Evaluation*, 12(10), 1–8.

- International Organization for Standardization (1986). ISO 8402-1986: Quality-Vocabulary. Retrieved from <https://www.iso.org/standard/15570.html>.
- Iphar, C., Napoli, A., & Ray, C. (2015). Detection of false AIS messages for the improvement of maritime situational awareness. In OCEANS'15 MTS/IEEE Washington (pp. 1–7).
- Kaczmarek, T., & Wećkowski, D. (2013). Harvesting deep web data through produser involvement. In Frameworks of IT Prosumption for Business Development (pp. 200–221). IGI Global.
- Kazemi, S., Abghari, S., Lavesson, N., Johnson, H., & Ryman, P. (2013). Open data for anomaly detection in maritime surveillance. *Expert Systems with Applications*, 40(14), 5719–5729.
- Kobus, J., & Westner, M. (2016). Ranking-type Delphi studies in IS research: step-by-step guide and analytical extension. In 9th IADIS International Conference (p. 28).
- Levy, Y., & Ellis, T. J. (2006). A systems approach to conduct an effective literature review in support of information systems research. *Informing Science*, 9, 181–211.
- Linstone, H. A., & Turoff, M. (2002). The Delphi method. *Techniques and Applications*, 53.
- Manyika, J., Chui, M., Groves, P., Farrell, D., Van Kuiken, S., & Doshi, E. A. (2013). Open Data: Unlocking Innovation and Performance with Liquid Information. Retrieved from [http://www.mckinsey.com/~media/McKinsey/dotcom/Insights/BusinessTechnology/Open data Unlocking innovation and performance with liquid information/MGI_Open_data_Full_report_Oct_2013.ashx](http://www.mckinsey.com/~media/McKinsey/dotcom/Insights/BusinessTechnology/Open%20data%20Unlocking%20innovation%20and%20performance%20with%20liquid%20information/MGI_Open_data_Full_report_Oct_2013.ashx).
- Małyszko, J., Abramowicz, W., & Stróżyna, M. (2016). Named Entity Disambiguation for Maritime-related Data Retrieved from Heterogenous Sources. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation*, 10(3), 465–477.
- Mendes, P. N., Mühleisen, H., & Bizer, C. (2012). Sieve: linked data quality assessment and fusion. In Proceedings of the 2012 Joint EDBT/ICDT Workshops on - EDBT-ICDT '12 (pp. 116–123). New York: ACM Press. doi:<https://doi.org/10.1145/2320765.2320803>
- MMO (2013). Spatial Trends in Shipping Activity. A report produced for the Marine Management Organisation, pp 46. MMO Project No: 1042.
- Neumaier, S., Umbrich, J., & Polleres, A. (2016). Automated quality assessment of metadata across open data portals. *Journal of Data and Information Quality*, 8(1), 1–29. <https://doi.org/10.1145/2964909>.
- Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: an example, design considerations and applications. *Information & Management*, 42(1), 15–29.
- Peter, B. (2010). Data quality. The key to interoperability. ISO Focus+, Volume 1, No. 2, February 2010, 28–29.
- Robey, D., & Markus, M. L. (1984). Rituals in information system design. *MIS Quarterly*, 5–15.
- Rogers, Y., Sharp, H., & Preece, J. (2011). Interaction design: beyond human-computer interaction. Wiley.
- Ruckhaus, E., Baldizán, O., & Vidal, M. E. (2013). Analyzing linked data quality with LiQuate. In Y. T. Demey & H. Panetto (Eds.), Confederated Int. Workshops on On the Move to Meaningful Internet Systems, OTM 2013 Workshops (Vol. 8186, pp. 629–638). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:<https://doi.org/10.1007/978-3-642-41033-8>
- Rula, A., & Zaveri, A. (2014). Methodology for assessment of linked data quality. In 1st Workshop on Linked Data Quality, LDQ 2014 (Vol. 1215). CEUR-WS. Retrieved from <http://www.scopus.com/inward/record.url?eid=s2.0-84908681022&partnerID=tZOtx3y1>
- Saaty, T. L. (1990). How to make a decision: the analytic hierarchy process. *European Journal of Operational Research*, 48(1), 9–26.
- Schmidt, R. C. (1997). Managing Delphi surveys using nonparametric statistical techniques. *Decision Sciences*, 28(3), 763–774.
- Stróżyna, M., Eiden, G., Filipiak, D., Małyszko, J., & Węcel K. (2016) A Methodology for quality-based selection of internet data sources in maritime domain. In: W. Abramowicz, R. Alt, B. Franczyk (Eds), Business Information Systems. BIS 2016. Lecture Notes in Business Information Processing, vol 255. Cham: Springer.
- Węcel, K., & Lewoniewski, W. (2015). *Modelling the quality of attributes in Wikipedia infoboxes. Lecture Notes in Business Information Processing* (Vol. 228). doi:https://doi.org/10.1007/978-3-319-26762-3_27
- Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for linked open data: a survey. *Semantic Web Journal*, 7(1), 63–93. <https://doi.org/10.3233/SW-150175>.