A Methodology for Quality-Based Selection of Internet Data Sources in Maritime Domain

Milena Stróżyna[1](✉), Gerd Eiden[2], Dominik Filipiak[1],
Jacek Małyszko[1], and Krzysztof Węcel[1]

[1] Poznań University of Economics and Business, Poznań, Poland
{milena.strozyna,dominik.filipiak,jacek.malyszko,
krzysztof.wecel}@kie.ue.poznan.pl
[2] LuxSpace Sarl, Betzdorf, Luxembourg
eiden@luxspace.lu

Abstract. The paper presents a methodology for identification, assessment and selection of internet data sources that shall be used to supplement existing internal data in a continuous manner. Several criteria are specified to help in the selection process. The proposed method is described based on an example of the system for the maritime surveillance purposes, originally developed within the SIMMO research project. As a result, we also present a ranking of concrete data sources. The presented methodology is universal and can be applied to other domains, where internet sources can offer additional data.

Keywords: Internet data sources · Quality assessment · Selection methodology

1 Introduction

Each information system need to be supplied with data in order to fulfil its functions. The data can come from various sources, depending on the system, its purposes and operating context. In systems used by organisations, sources of data can be internal (e.g. transactional data) or external, coming from the outside of an organisation (e.g. sensors, external systems and databases, Internet). Irrespective of the type of data used, each potential data source for a system needs to be appropriately defined and assessed. This procedure is crucial, while designing and developing a system [1].

The goal of this paper is to present a methodology for a selection of open internet sources that can be treated as an data source for an information system. Data from the Internet is used then to enhance data from other sources, such as legacy systems, sensors, internal databases, etc.

The general scope of this paper encompasses a procedure for identification, assessment and selection of internet data sources. The process of designing the proposed methodology was driven by the standard approach to the data quality, which defines quality as *"the totality of features and characteristics of a product*

W. Abramowicz et al. (Eds.): BIS 2016, LNBIP 255, pp. 15–27, 2016.
DOI: 10.1007/978-3-319-39426-8_2

or service that bears its ability to satisfy stated or implied needs" [2]. In case of information systems, these needs mean the functional and non-functional requirements. Therefore, we assumed that each potential data source for the system should be analysed and assessed taking into account two elements: (1) system's requirements; (2) a selected set of quality criteria. In the first step we select a set of quality measures, which are then used to assess data sources. Then, for each measure a rating scale and a weight is assigned. Finally, a method for calculating a quality grade and setting a selection threshold is specified.

The paper is built around the use case of an information system from the maritime domain, shortly described in the next section. Section 3 presents the related work in the area of data sources selection for the information systems. Then, in Sect. 4 a proposal of applied research methodology for identification and selection of internet data sources is presented. Section 5 describes the results of the project's work, where the proposed approach for sources selection was applied. In Sect. 6 we summarise the results.

2 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 the maritime security and safety by providing appropriate level of maritime surveillance. This, in turn, can be realised by providing tools supporting maritime stakeholders in analysis of the current situation at sea – creation of the so-called "Maritime Domain Awareness (MDA)". MDA implies collection, fusion and dissemination of huge amount of data, coming from many, often heterogeneous, sources. However, current capabilities to achieve this awareness are still improving and there is a need for development of dedicated information systems and tools. This need concerns especially systems, able to fuse in real-time data from various heterogeneous sources and sensors. To our best knowledge, currently there exists no maritime surveillance system which would automatically acquire and fuse AIS data with information available in internet sources. As a result, there is also no standard methodology for selecting and assessing the quality of internet sources to be used in such systems.

This challenge was addressed by the SIMMO project[1]. Within the project a system has been developed aiming at improving the maritime security and safety by providing high quality information about vessels and automatically detecting potential threats (i.e. suspicious vessels). The concept of the SIMMO assumes constant retrieval and fusion of data from two types of data sources, namely:

1. Satellite and terrestrial Automatic Identification System (AIS)[2], which provides inter alia information about location of ships and generic static information about them.

[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.

2. Open internet sources that provide additional information about ships, not included in AIS (e.g. flag, vessel type, owner).

In general, the SIMMO system integrates information from these two types of sources, what is essential for the better identification of vessels, and then detects suspicious vessels. For the efficient use of sources of the second type, a need to select appropriate sources emerged. The methodology for the quality assessment of potential internet sources had to be defined and adopted.

3 Related Work

3.1 Internet Sources Related to the Maritime Domain

The creation of the enhanced Maritime Domain Awareness and detection of suspicious vessels requires usage of different data sources. The sources that are applicable in the maritime surveillance domain can be divided into three categories.

The first and the most widely used are sensors. Sensors provide kinematic data for the observed objects in their coverage area and can be further divided into active (e.g. radar, sonar) and passive (which rely on data broadcasted intentionally by objects, e.g. AIS). A survey on sensors used in maritime surveillance can be found in [3].

The second category includes authorised databases, containing information about vessels, cargo, crew etc. However, most of them are classified and encompass inter alia port notifications sent by ships, HAZMAT reports, The West European Tanker Reporting System (WETREP), LRIT data centers, SafeSeaNet [4].

The first and the second category are basically accessible only to the maritime authorities, such as the coastguard. Therefore, 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 Web (hereinafter referred to as internet data sources). This data includes inter alia vessel traffic data, reports and news. More specifically, they can be divided as follows [5]:

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

The term open data refers to the idea of making data freely available to use, reuse or redistribute without any restriction [6]. In the maritime context, there are organisations and communities that provide their maritime related data online and make it accessible for the public. Examples are ports, publishing vessel traffic data as well as blogs, forums and social networks, which share information about maritime events [5].

Although there are various categories of data sources, in the existing maritime information systems usually only data received from sensors are used. The research in this area focuses on collection of sensor data, such as SAR, AIS, IR, video and radar data [3,7] or fusion of sensor and non-sensor data (for example inclusion of expert knowledge) [8–10]. The research, which additionally focuses on usage of open data for the purpose of maritime surveillance, is presented in [5].

3.2 Data Quality Assessment

There is no uniform definition of data quality nor a standard or a commonly used approach for assessment of data quality. ISO9000:2015 defines data quality as the degree to which a set of characteristics of data fulfils requirements [11]. In the information systems literature, a lot of various data quality attributes can be found. The examples are: completeness, accuracy, timeliness, precision, reliability, currency and relevancy [12]. Other such as accessibility and interpretability are also used. Wang et al. [13] identified nearly 200 such quality attributes. Still, no general agreement exists either on which set of dimensions define the quality of data or on the exact meaning of each dimension. Batini et al. [14] present different definitions of popular quality attributes provided in the literature.

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 quality of results for the purposes of the public statistics, described in the ESS Quality Assurance Framework [15]. In this standard, seven quality criteria were defined [16]: (1) *relevance* (the degree to which data meets 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).

When the quality attributes are defined, the next step is data quality assessment. Also in this matter, the literature provides a wide range of techniques to assess and improve the quality of data. In general, the assessment consists of several steps [14]: (1) *data analysis* (examination of data schemas, complete understanding of data and related architectural and management rules); (2) *data quality requirements analysis* (surveying the opinion 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 modelling* (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.

There exist a number of methodologies for quality assessment and quality measurement. Batini surveyed thirteen of them [14]. Nauman et al. [17] propose a quality driven source selection method using Data Envelopment Analysis technique. 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 a 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 involvement of the subject-matter experts (expert or heuristic techniques) [18].

4 Methodology

While designing and developing an information system, a key role plays the selection of data sources. These sources can be either internal or external (coming from outside the organisation). Irrespective of the type of data used, each potential data source needs to be identified and assessed. This procedure consists of several steps: (1) identification of potential data sources; (2) definition of quality criteria; (3) analysis of identified sources and assessment with regard to defined requirements (quality measurement); (4) selection of sources for a system. For selected sources a detailed design of data acquisition procedures takes place, including cooperation model (e.g. politeness policy). When data is obtained it has to be fused, i.e. a common data model meeting the initial system requirements has to be developed and used to organise new data in a homogeneous and integrated form. It entails semantic interoperability problems related to the interpretation of data coming from different sources. Although covered in the SIMMO project, the process of fusion data from various sources is a separate process from the source selection and as such is out of scope of this paper. In the following paragraphs, we describe in details the steps of the proposed method, using a use case from the SIMMO project. The method is presented in Fig. 1.

In order to identify, assess and select internet data sources for the SIMMO system and then to set up a cooperation model with the selected sources, a specific methodology has been followed. In the first step, potential data sources related to maritime surveillance were identified. To this end, conventional search engines (like Google) as well as meta search engines like Dogpile[3], Mamma[4], Webcrawler[5] were used. Apart from the search engines, also other data sources were analysed, including sources indicated in [19] and suggested by maritime experts, who were interviewed during requirements analysis for the SIMMO system. In this paper, we focused only on data sources which are used by the system regularly, meaning that they are constantly monitored for changes and

[3] http://www.dogpile.com/.

[4] https://mamma.com/.

[5] http://www.webcrawler.com/.

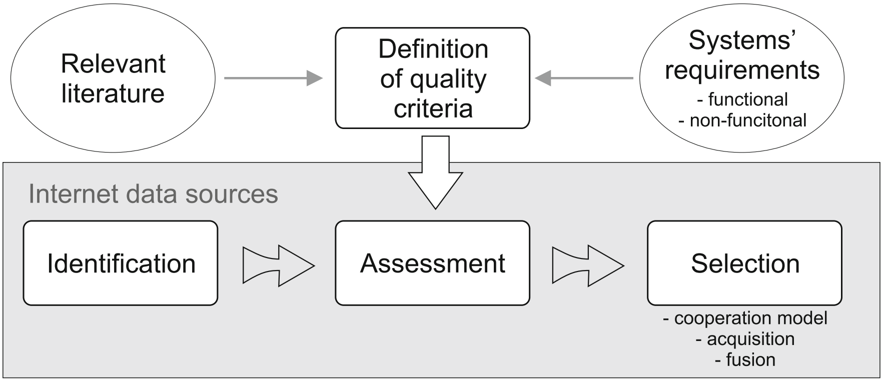

Fig. 1. Source selection method

data from them is retrieved at defined time intervals. The conducted analysis does not encompass the data sources which are to be used only once and will not be monitored for changes, e.g. internet source with a list of all ports worldwide.

In order to select sources of the highest quality and best suited to the SIMMO system, the identified data sources are assessed using specific quality criteria. For each identified data source, its features and characteristics are analysed and assessed taking into account the functional requirements defined for the SIMMO system and the selected set of quality criteria. Each potential source is assessed using the same set of quality measures.

In the proposed methodology, the data quality measures proposed by the European Statistical System (ESS) were adopted [15]. As a result, the following six quality measures are used: (1) Relevance (Usefulness); (2) Accuracy and Reliability (Completeness); (3) Timeliness and Punctuality; (4) Coherence and Comparability; (5) Accessibility (Availability); (6) Clarity (Transparency). These criteria can be adjusted to the specific of the developed system. For each measure a four-level rating scale is used (high = 3, medium = 2, low = 1, N/A = 0) and a weighting factor is assigned. After the assessment in each criteria, a final quality grade is calculated. The internet sources with the final mark above a defined threshold is then selected as a source for the system. In case of SIMMO, the threshold was set to 85 %. The weighting factors and quality threshold were assigned using the Delphi method [20].

In the final step, a cooperation model for each selected source is defined. The cooperation model should present how a cooperation with data provider (selected data source) will look like, including politeness policy and time intervals between data updates. For this end, each source had to be analysed with regard to existence of a defined politeness policy or terms of use.

5 Results

5.1 Identification of Internet Data Sources

As indicated in Sect. 4, the first step of the methodology for system's sources selection is the identification of potential sources. In the SIMMO case, potential data sources related to maritime surveillance and maritime domain were identified, using search engines, literature review and consultations with subject matter experts. As a result, over 60 different data sources available on the Web were found. The identified data sources are part of both the shallow and the deep Web. They provide 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 and login required (OR) – websites that provide information to users only after registration and logging (e.g. Equasis),
3. Data sources with partially paid access (PPA) – websites that after paying a fee provide a wider scope of information (e.g. MarineTraffic),
4. Commercial (paid) data sources (PA) – websites with only paid access to the data (fee or subscription required).

From all the identified sources, for further analysis we selected only the open data sources. At this stage, we eliminated the commercial data sources and websites that provide access to data only after prior authorization. The elimination of commercial websites results from the fact that they provide only very general, marketing information about data they have and access to any data is available only after paying a fee or signing a contract. Moreover, an attempt to make a contact with these data providers in order to get access to a sample data failed. Therefore, we have not been able to verify the data model or scope of data provided by these sources. As a result, only sources with a public content (open data sources) were selected for further analysis.

Similarly, two other data sources (IALA, SafeSeaNet) were sorted out due to the fact that access to the data requires application of the long-lasting procedure for the needed data with no guarantee that the access will be granted.

5.2 Internet Data Sources' Assessment

As a result of initial selection, 43 sources were assessed in details. As described in Sect. 4, each source was analysed from the point of view of six quality measures. Definitions of these criteria were adjusted to the specific of the SIMMO project (see Table 1). For each measure a weight was assigned, denoting an importance of a given measure in the final quality grade.

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 use-cases and 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: Own work.

Each source was assessed taking into account the following measures:

1. **Accessibility (A)** – here it is assessed, whether it is possible to retrieve data from a source using a crawler. It takes into account a structure of a source, technologies used in its development, a form in which data is published as well as source stability (changes of a structure, errors, unavailability of a service). This includes also such aspects like terms of use, privacy policy, a requirement for login or registration, access to data (fees, subscriptions) etc.,
2. **Relevance (R)** – what kind of information is provided by a source and whether this information matches the user requirements of the SIMMO system,
3. **Accuracy & Reliability (AR)** – it is assessed, whether information provided is reliable (a source (owner) of information is trusted). It evaluates also data scope (how much information is available), ship's coverage (information about how many ships is provided) as well as data accuracy (number of missing information),
4. **Clarity (C)** – it is assessed, whether a source provides appropriate description and explanation of data model and source of data (who is a data provider),
5. **Timeliness & Punctuality (TP)** – it is evaluated, how often data is updated (time interval between data availability and an event, which it describes) as well as what is a time delay in publishing updated information,
6. **Coherence & Comparability (CC)** – it is compared, whether the data provided in a source describes the same phenomenon or has the same unit of measure like data available in other sources.

The assessment of the identified sources was conducted by the SIMMO project's team, being experts and having experience in data retrieval from various data sources, including structured and unstructured internet sources. For this step, the Delphi method was utilized. For the selected sources, the experts assigned a grade in each criterion, using a four-level rating scale: high (grade 3), medium (grade 2), vlow (grade 1), N/A (grade 0).

The rate N/A means that an information required for a particular criterion (e.g. update interval or ship coverage) is not specified by a source and as a result, it was not possible to assess a source in this matter. In case of 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 source The results of quality assessment for each source is presented in Table 2.

5.3 Final Selection of Sources

After sources' assessment, the final selection took place. Firstly, all sources with Accessibility measure equals to *N/A* or *Low* were sorted out. This elimination results from the reasons indicated in the previous paragraph and the prohibition of using information from these sources.

With regard to data sources with *Low* Accessibility, this encompasses the sources with unstructured information (e.g. text in a natural language). We excluded them due to the fact that an automatic retrieval of this information would require a large amount of work and developing methods in the field of Natural Language Processing. As a result, we decided to include in the project only sources with structured or semi-structured information.

Secondly, sources with Relevance equal to *Low* were eliminated as well. It results from the fact that it's pointless to retrieve data that are not well-suited to the use-cases or requirements defined for the SIMMO system.

In the next step, a final quality grade for each source was calculated according to the formula:

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

where s means number of the analysed sources, $n = (1, 6)$, x_i means the grade assigned by the experts to a given quality measure i, and w_i means the measure's weight.

Based on the calculated quality grade, a ranking of sources was created (see Table 2). From the ranked list of sources, only sources with a final grade above a defined threshold were selected for usage in the SIMMO system (the bold Source Names in Table 2). The quality threshold was defined by the experts at the level of 85 %.

5.4 Model of Cooperation with Data Owners

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

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	98,33%	Yes
	US Coast Guard	O	H	H	H	H	M	H	98,33%	Yes
	Maritime mobile Access and Retrieval System (ITU MARS)	O	H	H	H	H	M	M	91,67%	Yes
	Maritime-connector	O	H	H	H	N/A	H	H	90,00%	Yes
	ShipFinder	O	H	M	M	H	M	L	73,33%	No
	AIS HUB	O	H	M	M	M	M	L	70,00%	No
	Equasis	OR	N/A	H	H	H	M	H	68,33%	No
	IMO GISIS	OR	L	H	H	N/A	M	H	68,33%	No
	Vessel finder	O	N/A	L	H	H	H	H	66,67%	No
	FleetMon	OR	N/A	M	H	H	M	H	66,67%	No
	Lloyd's Register Ship in Class	OR	M	H	M	L	M	M	65,00%	No
	ShipSpotting	OR	N/A	H	H	H	H	L	56,67%	No
	World Shipping Register	PPA	N/A	M	M	M	H	H	55,00%	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	55,00%	No
	Seaagent	PPA	N/A	L	H	N/A	L	N/A	33,33%	No
Weather	ICM Meteo	O	L	H	L	M	L	L	40,00%	No
	Meteooffice	O	L	H	L	M	L	L	40,00%	No
	Sailwx	O	N/A	L	L	M	L	M	33,33%	No
Classification of ships	**International Association of Classification Societies (IACS) Vessel in class**	O	H	H	H	H	H	H	100,00%	Yes
	American Bureau of Shipping (ABS)	O	H	H	H	M	M	M	88,33%	Yes
	International Association of Classification Societies (IACS) Transfer of Class	O	H	H	M	H	M	M	81,67%	No
	ClassNK	O	N/A	H	H	M	M	M	58,33%	No
	Leonardo Info	OR	N/A	H	H	N/A	M	H	58,33%	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	96,67%	Yes
	Tokyo Mou	O	H	H	H	H	L	H	96,67%	Yes
	Mediterranean MoU	O	H	H	H	H	L	H	96,67%	Yes
	Black Sea MoU	O	H	H	H	H	L	H	96,67%	Yes
	Government of Canada - Port State Control	O	H	H	M	H	H	H	90,00%	Yes
	Indian Ocean MoU	O	M	H	H	H	L	H	86,67%	Yes
	Riyadh MoU	O	M	H	H	H	L	M	80,00%	No
	Latin America Mou	O	M	M	H	H	L	M	78,33%	No
	Paris MoU	O	N/A	H	H	H	L	H	66,67%	No
	Abuja MoU	O	N/A	H	H	M	L	H	63,33%	No
	Caribbean MoU	O	N/A	H	H	N/A	L	H	56,67%	No
Maritime crimes	ICC Commercial Crime Services	PPA	M	H	L	H	L	M	60,00%	No
	Maritime Safety Information	O	L	H	L	M	L	H	53,33%	No
Tankers	**Q88.com**	PPA	H	H	H	M	M	M	88,33%	Yes
	Auke Visser's International Supertankers	O	L	M	H	L	M	M	63,33%	No
	International Association of Independent Tanker Owners	PA	-	-	-	-	-	-	-	No
Container ships	Containership-info	O	H	M	L	M	M	H	73,33%	No
Fishing vessels	Commission for the Conservation of Antarctic Marine Leaving Resources (CCAMLR)	O	H	H	L	H	H	M	73,33%	No
	International Convention for the Conservation of Atlantic Tunas (ICCAT)	O	H	H	L	N/A	H	H	70,00%	No
	Indian Ocean Tuna Commission (IOTC)	O	H	H	L	N/A	H	H	70,00%	No
	Western & Central Facific Fisheries Commission (WCPFC)	O	H	H	L	N/A	H	M	63,33%	No
	FAO Vessel Record Management Framework (FVRMF)	O	M	H	L	L	M	H	61,67%	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	80,00%	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; P - Paid access

- 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 by the SIMMO system,
- type of source – whether retrieved content is published in shallow or deep Web and in what form data is 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 the website administrators defined, if any, defining which parts of their Web servers should not be accessed by crawlers as well as indicating the number of seconds to delay between requests,
- re-visit approach – how often the SIMMO system will retrieve information from a source, i.e. the intervals between consecutive downloads from the source, taking into account the politeness policy, if defined.

To sum up, the application of the proposed sources' selection methodology in the SIMMO project allowed us to identify, assess and finally choose open internet data sources of the highest quality, which are about to supply the SIMMO system with the maritime data. However, it needs to be stressed that the whole assessment procedure did not focus on the quality of data available in a given source, rather than on the quality of the source itself. The aspects of data quality retrieved from the selected sources (e.g. data completeness, validity, consistency, ambiguity etc.) were dealt in the project at the later steps of the system development. Due to limited volume of this paper it is impossible to present them here.

6 Discussion and Conclusion

The goal of this paper was to propose a methodology for identification, assessment and selection of internet data sources, which are about to be a source of information for an information system. In a nutshell, the proposed methodology consists of 5 steps, starting from potential sources identification and ending with definition of cooperation model. It can be used in designing an information system in any domain which requires acquisition of data available in the Internet. In the paper, the method is described and evaluated based on the running example of the information system from the maritime domain.

The performed analysis gave us an overview on the scope of the data related to vessels and maritime domain which is available on the Web and can be freely used in the maritime information systems. Moreover, the conducted analysis revealed that there is plenty of data sources with valuable information that unfortunately cannot be used due to strict terms of use or policies regarding prohibition on the use of any techniques for automatic retrieval of data published by a given source. There are also sources, which require prior written authorization to use their data. However, an attempt to get such authorization failed (there was no response from the information provider regarding our request for access) or the whole procedure is long-lasting and requires engagement of public authorities.

At the moment, the proposed method has not been validated in other domains or industries. Nevertheless, we believe that it could be utilized for assessment of potential internet sources for traffic monitoring systems used in other transportation areas, such as railway, road or air. Analysis of possible exploitation in these domains may be the subject of the future work on the proposed methodology.

Moreover, the future work may encompass proposing additional analysis steps, which would focus more on the quality of the data itself (not only on the quality of the source). It would require specification of additional quality measures and development of method(s) for automatic assessment of the data quality as soon as the data is acquired. Also inclusion of additional attributes, e.g. domain-related, for assessing the source quality may also be considered.

Acknowledgements. This work was supported by a grant provided for the project *SIMMO: System for Intelligent Maritime MOnitoring* (contract no. A-1341-RT-GP), financed by the Contributing Members of the JIP-ICET 2 Programme and supervised by the European Defence Agency.

References

1. Robey, D., Markus, M.L.: Rituals in information system design. MIS Q. **8**, 5–15 (1984)
2. International Organization for Standardization: ISO 8402–1986 (GB/T6583-1992): Quality-Vocabulary, June 1986
3. Vespe, M., Sciotti, M., Battistello, G.: Multi-sensor autonomous tracking for maritime surveillance. In: International Conference on Radar, 2008, pp. 525–530. IEEE (2008)
4. European Commission: Integrated Maritime Policy for the EU. Working document III on Maritime Surveillance Systems (2008)
5. Kazemi, S., Abghari, S., Lavesson, N., Johnson, H., Ryman, P.: Open data for anomaly detection in maritime surveillance. Expert Syst. Appl. **40**(14), 5719–5729 (2013)
6. Alonso, J., Ambur, O., Amutio, M.A., Azañón, O., Bennett, D., Flagg, R., McAllister, D., Novak, K., Rush, S., Sheridan, J.: Improving access to government through better use of the web. World Wide Web Consortium (2009)
7. Rhodes, B.J., Bomberger, N.A., Seibert, M., Waxman, A.M.: Maritime situation monitoring and awareness using learning mechanisms. In: Military Communications Conference, MILCOM 2005, pp. 646–652. IEEE (2005)
8. Fooladvandi, F., Brax, C., Gustavsson, P., Fredin, M.: Signature-based activity detection based on Bayesian networks acquired from expert knowledge. In: 12th International Conference on Information Fusion, FUSION 2009, pp. 436–443. IEEE (2009)
9. Riveiro, M., Falkman, G., Ziemke, T.: Improving maritime anomaly detection and situation awareness through interactive visualization. In: 11th International Conference on Information Fusion, 2008, pp. 1–8. IEEE (2008)

10. Helldin, T., Riveiro, M.: Explanation methods for Bayesian networks: review and application to a maritime scenario. In: Proceedings of the 3rd Annual Skövde Workshop on Information Fusion Topics (SWIFT 2009), pp. 11–16 (2009)
11. Peter, B.: Data quality. The key to interoperability (2010)
12. Wang, R.Y., Reddy, M.P., Kon, H.B.: Toward quality data: an attribute-based approach. Decis. Support Syst. **13**(3), 349–372 (1995)
13. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. J. Manage. Inf. Syst. **12**, 5–33 (1996)
14. Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for data quality assessment and improvement. ACM Comput. Surv. **41**(3), 16:1–16:52 (2009)
15. European Statistical System: ESS handbook for quality reports. Eurostat (2014)
16. European Parliament: Regulation (EC) No 223/2009 of the European Parliament and the Council of 11 on European statistics and repealing Regulation (EC, Euratom). Official J. Eur. Union **52** (2009)
17. Naumann, F., Freytag, J.C., Spiliopoulou, M.: Quality-driven source selection using data envelopment analysis. In: Proceedings of the 3rd Conference on Information Quality (IQ), Cambridge, MA (1998)
18. Dorofeyuk, A., Pokrovskaya, I., Chernyavkii, A.: Expert methods to analyze and perfect management systems. Autom. Remote Control **65**(10), 1675–1688 (2004)
19. Kazemi, S., Abghari, S., Lavesson, N., Johnson, H., Ryman, P.: Open data for anomaly detection in maritime surveillance. Expert Syst. Appl. **40**(14), 5719–5729 (2013)
20. Brown, B.B.: Delphi process: a methodology used for the elicitation of opinions of experts. Technical report, DTIC Document (1968)