TRANSNAV
http://www.transnav.eu
the International Journal
on Marine Navigation
and Safety of Sea Transportation
Volume 10
Number 3
September 2016
DOI: 10.12716/1001.10.03.12

Named Entity Disambiguation for Maritime-related Data Retrieved from Heterogenous Sources

J. Małyszko, W. Abramowicz & M. Stróżyna
Poznań University of Economics and Business, Poznań, Poland

ABSTRACT: The article concerns integration and disambiguation of data related to the maritime domain. A developed system is described, which collects and merges data about several maritime-related entities (vessels, vessel types, ports, companies etc.) retrieved from different internet sources and feeds the data into a single database. This process is however not trivial. There are few challenges, which need to be faced to successfully conduct it. Firstly, in different sources, entities may be referenced to in different ways, for example, by using different text strings. Additionally, some of these references may be ambiguous, i.e. potentially the reference may point to more than one entity. To enable efficient analysis of data coming from different sources, such ambiguities must be resolved automatically as a preprocessing step, before the data is uploaded to the database and utilized in further computations. The aim of the disambiguation process is to assign artificial, unique identifiers to each entity and then, if possible, automatically assign these identifiers to each data item related to a given entity. In the article, developed methods for resolving such ambiguities are discussed and their evaluation is presented.

1 INTRODUCTION

Maritime Surveillance is an essential priority for World's Security, both for the civilian as well as for the military sector. In this aspect, Maritime Domain Awareness (MDA) plays a critical role. MDA is "the effective understanding of any activity associated with the maritime environment that could impact upon the security, safety, economy or environment" [1]. Thus MDA implies the collection, fusion and dissemination of enormous quantities of data in order to build intelligence and create a comprehensive Common Operating Picture (COP). According to [2], MDA is the sine qua non of maritime security and depends on surveillance and exchange of information between the international communities. However, current capabilities to achieve that awareness are still under development, what especially concerns the integration of data from different sources and increase of the quality of maritime-related data. Therefore, the current potential stemming from utilization of this data is not yet fully exploited, particularly in view of data fusion and the use of intelligent data analysis tools.

To fulfil this potential, methods and systems for creating a complete maritime situation picture are required. This includes for examples systems, which integrate static and dynamic data about vessels from AIS with information from external sources (further called as ancillary information). Such systems would support operators in charge in the process of monitoring and controling of the maritime traffic as well as in the OODA loop [3]:

– Observe: to know what is going on,
– Orient: to understand what is going on,

– Decide: to weight the options and their impact,
– Act: to carry out the decision.

In this article, we describe part of the work that was conducted during SIMMO (System for Intelligent Maritime MOnitoring) project, aiming at integration of data from multiple sources to enhance MDA. The main goal of the SIMMO was to develop a prototype of a system, based on the state-of-the-art information fusion and intelligence analysis techniques, which generates an enhanced Recognised Maritime Picture (RMP) and thus support a user in situation analysis and decision-making. This aim was addressed by providing information about vessels of higher quality and by automatic detection of potential threats (suspicious vessels) with regard to defined criteria. The system is addressed to different stakeholders and entities from the maritime domain.

As mentioned above, the SIMMO system collects and fuses data from two types of data sources: AIS and selected internet data sources. In this article we focus only on the latter type. Having data retrieved from selected internet sources (which is retrieved by web scrappers), the system performs merging and integration of this data into a consistent data set. The data itself concerns different maritime-related entities, such vessels, flags, ports, vessel types, classification societies and companies. Each of these entities is described in more detail in section 3

The data integration is a complex process and there are few challenges, which need to be faced to successfully conduct it. Firstly, in each data source the same entity may be referenced in different ways and different categories may be used to describe the same issues. For example, different words (names) may be used to call the same entity (e.g. a port or a ship) or categories used in two sources may be developed on different levels of granularity. Therefore, before the data is added to the database, such differences must be recognized and the data needs to be aligned. For example, the system should recognize which entities are being referenced to in the data and, based on that, assign to this data an unique identifier, which can be easily used for subsequent analysis. This process is called disambiguation. In this article we present a set of methods, designed and implemented within the SIMMO project, which aim at solving such issues.

The article is organized as follows. First, a brief analysis of related work is described to give the reader the context of the task, which had to be performed to reach the objectives of the research. Next section 3 constitutes the main part of the article and contains description of the developed approaches for all analyzed entities: vessels (subsection 3.1), flags (3.2), ports (3.3), vessel types (3.4), classification societies 3.5 and companies (3.6). The article is concluded with a short summary and an outlook on a possible future research directions.

2 RELATED WORK

2.1 *Disambiguation process*

The research is related to ETL (Extract, Transform, Load) task. ETL refers to a process in a database usage, especially in a data warehouse, that:

– extracts data from homogeneous or heterogeneous data sources; in ETL, these are usually databases, which may be accessed directly or using dedicated API. In traditional ETL research, an important issue is reducing an overload of the data source, resulting from extracting data from it (to ensure, that the performance of the original data source will not suffer) and, at the same time, keeping the data as up-to-date as possible [4]. Still, in case when the sources are web pages (as is the case in SIMMO) this issue is no longer valid, as the rate on which queries may be sent to a web page is strictly defined.
– transforms the data in order to store it in the proper format or structure, for the purposes of querying and analysis; transformation steps used here often are ad-hoc, developed to fit a given situation, and straightforward if studied individually. Still, as the number of such transformation steps of this kind may grow, a proper approach should be utilized to ensure proper efficiency and elegance in terms of semantics [4].
– loads the data into the final target (database or data warehouse) for possible exploitation.

The area of interest of this article is distinguished from traditional ETL as it focuses more on the issue of merging data from different sources than on the whole process. One of the most important tasks in the area of data integration, which was conducted during our work, is entity disambiguation, which in the literature is also referred to as duplicate detection, record linkage, reference matching or entity-name clustering and matching problem [6]. It is a well-known problem in the area of data integration. It results from the fact that references to a single entity may be different due to different reasons, such as typographical errors, abbreviations etc. [6]. The mentioned problem is especially important to handle when data from many different sources is to be integrated. As different systems are developed and maintained by different parties, often to serve specific needs, in these systems the same entities may be referenced in completely different way [7].

According to [7], the following steps should be followed to perform the discussed task:
1 data analysis, which goal is to identify errors and inconsistencies that need to be removed,
2 definition of transformation workflow and mapping rules, which as a result is to provide methods and their implementations for data disambiguation,
3 verification, which goal is to evaluate to what extent the methods developed in the previous step give expected results; this step, together with the previous one, may be performed iteratively multiple times,
4 transformation, which processes the available data using methods selected during previous steps and updates the available database with final values,
5 backflow of cleaned data, which is updating the source database with the new, cleaned data (if possible).

Basic tools used for entity disambiguation problem are string similarity measures. These measures, having at input two strings, return a numerical value representing distance (or similarity) between them.

Based on such measures, for example, two strings which were found to be very similar to each other may be recognized as referring to the same entity (the difference between them may result, for example, from misspelling [8]). A simple, well known string similarity measures are Levenshtein distance [6] and Jaro distance [9].

Using the string similarity measures on attributes used to identify entities, it is possible to match the strings based on similarities between the values of these fields. Still, even greater challenge must be faced when there is no single uniquely identifying field for a certain entity. In such situations, multiple fields must be compared to establish some similarity measure between the two records [10].

To successfully perform entity disambiguation, lexical resources may also be needed to identify different ways, how a certain entity may be referenced to. For example, in paper [11] one of resources that was used for entity disambiguation was Disambiguation Dictionary that maps all ambiguous proper names to the set of unique entities they refer to. An example given in the mentioned article is similar to many cases which were encountered also in the SIMMO. Let's assume a situation when an abbreviation ACC is used, which refers to an entity, which for the system is known under a main name e.g. American College of Cardiology. Such mapping cannot be easily identified using, for example, Jaro measure. This problem may be solved if there is a proper dictionary, in which alternative names for known entities are defined. Additionally, in many situations such mappings may be ambiguous: e.g. ACC may also be Asian Cricket Council. To resolve such difficulties, usually additional data must be taken into account, e.g. context, in which a given word appears.

2.2 *Maritime-related internet data sources*

As it was indicated in the Introduction, creation of the enhanced Maritime Picture requires usage of different data sources. The data sources, which are applicable in the maritime surveillance domain, can be divided into three categories. The first and the most widely used are sensors, which include kinematic data for the observed objects in their coverage area and can be further divided on active (e.g. radar, sonar) and passive (which rely on data broadcasted intentionally by objects, e.g. AIS, LRIT). A survey on sensors used in maritime surveillance can be found in [12].

The first and the second category of sources are basically accessible only to the maritime authorities. Therefore they can be referred as closed data sources. Moreover, most of them do not publish data in any way on the Internet.

The third category consists of data sources, which are publicly available via Internet (hereinafter referred to as internet data sources). This data includes inter alia vessel traffic data, reports and news. There are organizations and communities that provide the maritime-related data online and make it accessible for the public. For example, there are different organizations, such as ports, that publish their vessel traffic data or their facilities information online. In addition, there are various online communities such as blogs, forums and social networks, which provide the possibility of sharing information about maritime events [13]. The main advantages of using such internet data sources are: possibility to reveal facts, which are not reported to the maritime authorities or available in their databases, global context of data and lack of legitimate limitations of exchanging data between different countries.

The maritime-related internet sources can be also divided into shallow and deep sources. The former are soyrces, which are indexable by conventional search engines, like Google or Yahoo. The deep sources consist of online databases that are accessible via Web interface, but poorly indexed by regular search engines and, in consequence, not available through regular Web search [14]. Such web pages are not directly accessible through static URL links, but rather dynamically generated as response to queries submitted through the query interface of an underlying database [15].

The deep web is an important source of information in the maritime domain. The analysis conducted within the SIMMO project revealed that there is a number of online databases, containing valuable information on various maritime entities, such as vessels, ports, ship owners etc.

As a result, there are different kind of data sources in the maritime domain that provide heterogeneous data regarding maritime entities. However, in the existing maritime surveillance systems, usually only the data received from sensors are used [16, 12]. Non-sensor data includes for example expert knowledge, which is further fused with sensor data [17]. Mano et.al. [18] proposed a system that collects data from radars and databases such as environmental database, Lloyd's Insurance and TF2000 Vessel DB. Ding et. al. [19] in turn proposed an architecture of a centralized integrated maritime surveillance system for the Canadian coasts, fusing HFSWR, ADS (Automatic Dependant Surveillance) reports, visual reports, information sources and radar. The solely research, which focuses on usage of open data available on the Internet for the purpose of maritime surveillance, is presented in [13].

3 RESULTS

In the SIMMO system, data about different maritime-related entities is retrieved from several internet sources and then combined into a single data model. These entities are:

- vessels,
- ports, which may be referenced to in many different contexts, for example current destination of a given vessel, home port for a vessel, location where vessel inspections takes place etc.,
- flags, corresponding to the country of registration of a given vessel,
- classification societies, which are organizations providing classification and statutory services and assistance to the maritime industry, as well as regulatory bodies with regards to maritime safety and pollution prevention, based on the

accumulation of maritime knowledge and technology[1],
– companies, which may be in certain relationships with vessels (e.g. owners or managers).

It is crucial to ensure that, as a result of data integration, it is possible to easily identify, which entity a particular data item refers to, regardless the source from which it was retrieved. To be able to do that, in the data model artificial identifiers have been introduced which are assigned to all entities. Such identifiers have the following characteristics:
– data items concerning the same entity should have the same ID assigned,
– data items concerning different entities should have different IDs assigned.

Having such IDs assigned, it is possible to query the database using regular SQL queries and retrieve required results, regardless the fact that in different data sources the same entity may be referenced to in different ways. Still, the main challenge is how to automatically assign the identifiers to the entities to ensure that the two characteristics of IDs described above are satisfied to the greatest possible extent. The process of assigning IDs to different entities is called entity disambiguation.

3.1 *Vessels*

In the SIMMO project, the main focus is put on data about vessels. Thus, apart from the disambiguation process, an additional step is performed in the system, aiming at fusing data into a single record. The fusion is understood as choosing one, final value for each attribute of a given vessel which is then used by the analytical module and presented to a system's user in the display module. Based on the data fusion, a single record with values for all ship's attributes is generated. This record, based on a set of defined rules, is most likely to be correct and valid. Below, in points 3.1 and 3.1 the process of disambiguation and fusion of vessel data is presented.

Vessel data disambiguation. The vessel data disambiguation is a process of assigning the same identifier to each data record concerning the same vessel (such identifier should be unique to a given vessel). A schema representing an example of a vessel disambiguation is presented in Figure 1, where are two records with selected data about static vessel features from two different sources (MarineTraffic and Maritime Connector). Let us assume that it was noticed that call sign and vessel name in both records are equal. Based on that it may be decided that these records concern the same vessel. In such situation, to both records shipId is assigned. Also, this shipId is to be assigned to any other data item which concerns this particular vessel.

In the research it was assumed that disambiguation of vessel data may be performed similarly as it was done in the example above, that is by checking if values of a certain attribute (or a collection of attributes, e.g. pairs of attributes, as in example above) in records coming from different sources are equal. As soon as the system identifies that there is a match between values of some attributes, the same shipId is to be assigned to both records. To ensure that such processing will give as good results as possible, it is important to correctly define in what order different attributes will be analyzed in search for the match. For example, first the attribute should be analysed, which is believed to give the most reliable results and if the match cannot be found (e.g. because values of such attributes may be missing), it should be moved to less reliable ones.

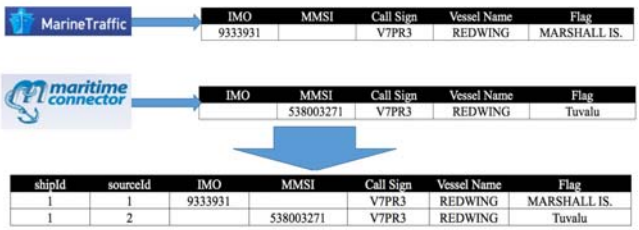

Figure 1. A simple schema presenting the goal of the vessel data merging process.

Vessels are characterized by a number of different attributes, which may be used for disambiguation purposes. Some of these attributes are specifically assigned by various organizations to enable unique identification of vessels in certain contexts. In the analyzed data sources these attributes are:

– IMO (International Maritime Organization Ship Identification Number Scheme) – numbers assigned permanently to each ship for identification purposes. That number should remain unchanged upon transfer of the ship to other flag(s) and is inserted in the ship's certificates[2],
– MMSI (Maritime Mobile Service Identitiy) – nine digit number used by several systems (including AIS) to uniquely identify a ship or a coast radio station. MMSIs are regulated and managed internationally by the International Telecommunications Union in Geneva, Switzerland[3],
– Call Sign,
– Vessel Name.

[1] http://www.iacs.org.uk/document/public/explained/Class_What Why&How.PDF, accessed 2016-03-23

[2] http://www.imo.org/en/OurWork/MSAS/Pages/IMO-identification-number-scheme.aspx, accessed 2016-04-01

[3] http://www.navcen.uscg.gov/index.php?pageName=mtMmsi, accessed on 2016-04-01

Table 1. Duplicates in source attribute value pairs for attributes, which potentially may be used for disambiguation of vessel data records

Attribute	# of duplicates (all values)	% of duplicates (all values)	# of duplicates (distinct values)	% of duplicates (distinct values)
IMO	361	0.098	178	0.088
MMSI	2043	0.559	987	0.321
Ship Name + Call Sign	2268	1.010	1100	0.541
Call Sign	26109	11.630	5261	2.979
Ship Name	157150	24.988	42230	11.433

Some of these identifiers are assigned by international organizations and are to be unique on the worldwide scale (e.g. MMSI and IMO). Thus, if two records from two distinct sources share the same MMSI or IMO number, it is highly probable that they concern the same vessel. For other attributes such assumption can be used with less certainty, as they may not be distinct.

For the above listed attributes it is estimated to what degree a given attribute is reliable as a unique identifier of a vessel. Such reliability is estimated based on the following heuristics. For each attribute it is checked how many times its value is duplicated in a single data source. For example, for MMSI it is counted how many times its value is duplicated in MarineTraffic, next how many times the value is duplicated in Maritime Connector etc. In the end, these numbers are summed. The more duplicated values are found, the less reliable this attribute is as far as unique identification of vessels is concerned. Thus, the attributes are ordered in a descending manner according to the number of such duplicates (relatively to the number of all values of such attribute in the database). The ordered list is then used in disambiguation process, i.e. the disambiguation is performed using in the first place the most reliable attributes.

The results of the data analysis are presented in table 1. Second and third column refer to all available data records (i.e. in how many records there are values, which are duplicated), while fourth and fifth column refer to distinct values (e.g. if distinct values of a given attribute are analysed, how many of them appear more than once in a single data source). Based on the results, the attribute with the highest reliability is IMO number, as duplicates occur in less than one per mille of cases. For MMSI, duplicates occur in more than a half per cent of cases, what still may be considered as reasonably low. Therefore it also can be used in disambiguation process. Still, for Call Sign and Ship Name, duplicates are much more common and thus disambiguation based on these attributes is likely to give much worse results. However, they may be used together, since duplicates for combination of Ship Name and Call Sign occur in about 1% of cases.

Another important issue concerning different (sets of) attributes is how often a certain value of the attribute appears in two or more data sources. Obviously, only if the value appears in more than one data source, it may be used to identify that two records from different data sources refer to the same ship. In table 2 statistics concerning this issue are presented.

Finally, table 3 presents statistics on how many rows are affected if the disambiguation is performed using the described approach, in the order presented in the table (first based on IMO, then on MMSI etc). The value in the second column takes into account the fact that disambiguation was already performed based on previous attributes. Thus, if the row was disambiguated based on IMO, it is not further analyzed whether it can be disambiguated also based on MMSI. The data in the last row reflects for how many rows there were no matches for the analysed attributes (or the attribute set). The third column contains values from the second column divided by the number of all rows with data about vessels and sums up to 100.

Table 2. How often the same value of different attributes may be found in more than one source (Ship Name + Call Sign row works on pairs of values of these two attributes)

Attribute	# of distinct values	% of distinct values
IMO	102515	50.627
MMSI	48150	15.700
Ship Name + Call Sign	17334	8.569
Call Sign	22124	12.884
Ship Name	79407	22.677

Table 3. How many rows are affected when merging of vessel data is conducted in the order presented in the table, from top to the bottom (Ship Name + Call Sign row works on pairs of values of these two attributes)

Attribute	# of rows affected	% of rows affected
IMO	269662	42.518
MMSI	29744	4.690
Ship Name + Call Sign	678	0.107
Call Sign	1405	0.222
Ship Name	25047	3.949
Not merged	308371	48.621

In our research, the vessel data disambiguation was performed according to the described approach and its results are presented in table 3. The attributes Ship Name and Call Sign were skipped as they were the least reliable attributes.

The proposed approach could be further extended by using additional attributes of vessels in the disambiguation process. In the database, many additional attributes of vessels were collected, such as flag, length, year of built and owner. These attributes may be used as an extra disambiguation information for data, for which less reliable attributes were utilized in the disambiguation, such as Ship Name (for example, both Ship Name and flag attributes must be equal to decide that both rows concern the

same vessel). Still, as the number of rows, which could be disambiguated based on Call Sign and Ship Name was relatively low, this issue was skipped in the performed work.

Vessel data fusion. The goal of data fusion is to select for each attribute describing a certain vessel and from all records describing that vessel, a single attribute value which will be considered to be the most accurate one. For example, let's assume that we have three records from three different sources for vessel with shipId = 1. According to source A, the flag for this ship is Poland, according to B it is Germany, and to source C again it is Poland. The goal of data fusion is to select one of these values, Germany or Poland, to be the primary value for this attribute. The record with fused data is to consist of such primary values for each attribute, as presented on picture 2.

The data fusion may be performed based on:
- selecting the most common value, i.e. the value that occurs in the data sources most often. It may be assumed that the value is correct because many or most sources report exactly the same value (Argumentum ad populum-like inference),
- assigning different priorities to different data sources (based on some previous assessment of the data sources) and selecting the value from source with the highest priority. The priority should reflect how reliable the source is according to the conducted assessment,
- analysis of agreement between different attributes. For example, first signs of a Call Sign correspond to the flag of the vessel. Thus, if the value of Flag attribute is different than what was expected from the Call Sign, this value may be chosen to be treated as less reliable one.

shipId	sourceId	IMO	MMSI	CallSign	VesselName	Flag
1	1	9333931		V7PR3	REDWING	MARSHALL IS.
1	2		53800327	V7PR3	REDWING	Tuvalu
1	3	9333931			REDWING	Tuvalu

shipId	IMO	MMSI	CallSign	VesselName	Flag
1	9333931	53800327	V7PR3	REDWING	Tuvalu

Figure 2. A simple schema of a vessel data fusion.

The might be situations, when a given the value of a given attribute is provided only in one source. In this case, this value is to be used in the fused record. Also, in many cases a given attribute will have the same value assigned in each record concerning a given ship (i.e. many sources provide the same value of a given attribute). In such situation this value is, obviously, going to be chosen as the primary value. In other cases, if there are different values assigned in records from different sources (i.e. different data sources provide different values of a certain attribute), part of the values must be discarded and only the one that is chosen as the primary value is put in the final, fused record. In the developed system, for each vessel's attribute a rule on how its values are fused was chosen by an expert.

3.2 *Flags*

In the data sources used in the research, each vessel has the flag assigned, which reflects its country of registration. A flag is referred by a string being a name of a given country. Although each country has exactly one name, there may be different variants how the name is written, e.g. due to abbreviations of country names or spelling errors. For example, one can easily find different ways how United States of America is referred to in different data sources:
- USA (US)
- U.S.A.
- United States of America
- United States
- UnitedStates (without space)

Apart from that, sometimes the flag name does not refer to the name of the country, but to one of its territories, e.g. Isle of Man and not the United Kingdom. In some scenarios, it might be useful to recognize, based on the name of the territory, what is the main country associated with a given string.


Mexico–United States relations refers to the foreign relations between the United Mexican States (*Estados Unidos Mexicanos*) and the United States of America. The two countries share a maritime and land border in North America. Several treaties have been concluded between the two nations bilaterally, such as the Gadsden Purchase, and multilaterally, such as the North American Free Trade Agreement. Both are members of various international organizations, including the Organization of American States and the United Nations.

```
<p><b>Mexico–United States relations</b> refers to the <a
href="/wiki/Foreign_relations" title="Foreign relations">foreign
relations</a> between the <a href="/wiki/Mexico"
title="Mexico">United Mexican States</a> (<i>Estados Unidos
Mexicanos</i>) and the <a href="/wiki/United_States" title="Unite
States">United States of America</a>. The two countries share a
maritime and land <a
href="/wiki/Mexico%E2%80%93United_States_border" title="Mexico–
United States border">border</a> in <a href="/wiki/North_America"
title="North America">North America</a>. Several treaties have
been concluded between the two nations bilaterally, such as the <a
href="/wiki/Gadsden_Purchase" title="Gadsden Purchase">Gadsden
Purchase</a>, and multilaterally, such as the <a
```

Figure 3. A paragraph (together with a part of its HTML code) from Wikipedia article about Mexico – United States relations, in which phrase "United Mexican States" is used as an anchor to link to the article, which name is Mexico. Based on such links, Wikipedia lexicalization dataset is generated

In the developed system, the list of flags to be used was defined upfront. For each flag a single string was assigned as its main name and a numerical identifier was assigned as well, called flagId. The goal of flag disambiguation is to, for a certain string representing a flag's name , identify which flagId this string corresponds to.

A basic prerequisite to perform such identification is a comprehensive lexical resource, containing for example flag name variant flagId mappings. Having a string representing a certain flag , the system can check in the lexical resource whether there is a mapping with a given flag name variant and, based on that, assign appropriate flagId (the one that is flagId paired with the given flag's name variant). Obviously, the crucial factor to enable successful disambiguation of flags is to obtain a comprehensive list of such flag name variant flagId mappings. First of all, the system should be able to assign flagId to any flag name variant found in the corpus. This can be done relatively easily, as the number of unique flag name strings in the available corpus is less than 600. In this case, a human expert is able to manually analyze all the cases and add necessary mappings to

ensure a 100% of coverage of flag name variants from the corpus in the lexical resource.

Still, the aim of the research was to develop a method which would allow to obtain also other mappings, not available in the initial corpus and which would extend the lexicon. The extended lexicon would be necessary in case when a new data (e.g. data from a new data source) is added to the system, containing previously unknown variants of a flag's name.

The resource which was used to automatically generate such extended lexicon was lexicalization dataset from DBpedia project. "DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web"[4]. The mentioned lexicalization dataset may be understood as a lexicon containing a list of Wikipedia concepts (i.e. article names) and their alternative names (i.e. text strings which may be used to refer to these concepts).

The lexicalization dataset is generated automatically based on analysis of hyperlinks within Wikipedia (the so-called interwiki links). In mamy cases, the Wikipedia article includes links, which point to some other Wikipedia pages. In such links, often the text of an anchor can be treated as an alternative name to the concept the link points to. We refer to such anchor texts as surface forms. As an example please refer to Figure 3, based on which phrase (surface form) "United Mexican States" may be identified to be an alternative name of the concept "Mexico". Thus, if it is possible to figure out that a given Wikipedia concept corresponds to a certain flag, all surface forms of links pointing to that concept can be automatically got and these surface forms can be considered as the alternative flag names.

The lexicalization dataset is provided by DBpedia in a form of a plain text file in a defined format. A sample of data from this file is presented in Listing 1.

```
<http://dbpedia.org/resource/Poland>   <http://lexvo.org/
ontology#label>     "Poland"@en     <http://dbepdia.org/
spotlight/id/Poland---Poland> .
<http://dbpedia.org/resource/Poland>   <http://lexvo.org/
ontology#label>     "Polish"@en     <http://dbepdia.org/
spotlight/id/Poland---Polish> .
<http://dbpedia.org/resource/Poland>   <http://lexvo.org/
ontology#label>   "Republic   of   Poland"@en   <http://
dbepdia.org/spotlight/id/Poland---Republic_of_Poland>.
```

Listing 1. Selected lines from DBpedia lexicalization dataset with surface forms pointing to the concept "Poland"

The surface forms, which point to the concept name are often correct alternative names of a given country. However, in some cases it may turned out, that some of the retrieved surface forms are useless from the point of view of flags disambiguation process and only introduce the noise. For example, the surface forms pointing to the concept Poland include "Polish", "Poland's", "Polish-born" and "Pole". These surface forms are unnecessary, as flag's names in data sources are referred to using nouns.

Such unnecessary variants may be easily filtered out by discarding words ending with a predefined sequences (e.g. "ish", "-born" and "'s").

[4] http://wiki.dbpedia.org/

The above-described inference on flag name variants is correct only when it is possible to connect the name of the flag, as known to our system, with the name of the Wikipedia concept corresponding to the given country. This inference process can be done in an automatic manner only when the flag name and name of the Wikipedia concept are exactly the same. Additionally, it must be ensured that the found concept indeed concerns the given country and is not other concept with the same name. For this purpose SPARQL query is used[5] (correctly defined SPARQL query may ensure that a given concept indeed is a country). In case, when the correct concept cannot be found in this way, the additional processing must be conducted, based on the following procedure:

1 Take the main flag name as known to the system and check, if there are some interwiki links in the DBpedia with surface forms equal to this string. Fetch the list of the matching surface forms and concepts, which these surface forms point to.
2 Fetch from the DBpedia a list of all concepts which refer to existing countries using a SPARQL query.
3 Make an intersection of two sets obtained in the step 1 and 2 and based on that identify the name of the concept corresponding to a given country.
4 Get surface forms of all interwiki links pointing to the found concept and add them as flag name variants.

Using the above-described approach, in total it was possible to extract around 1500 flag name variant – flagId pairs. Thanks to that, the developed system was able to automatically disambiguate flag name variant for almost every flag name string which was retrieve from internet data sources. For the remaining flag names, which still could not be disambiguated, appropriate mappings were added manually by experts to ensure full coverage. Finally, the developed solution was evaluated manually by an expert, who was shown a sample of 300 flag name strings as found in data sources together with flagIds assigned by the system. According to the expert, the system performed the disambiguation correctly in 299 out of these 300 cases (more than 99.6%).

3.3 *Ports*

In the used internet data sources, ports are used in different contexts, including:

- home port of a particular vessel,
- port visited by a ship,
- current vessel destination,
- port in which the Port State Control inspection took place.

Similarly to the flags, in the SIMMO system there is a predefined list of known ports and each port has a unique identifier assigned the portId (which is an integer value) as well as the main name of the port (a text string). Additionally, for each port its location and LOCODE[6] are specified.

[5] SPARQL is a semantic query language for databases, able to retrieve and manipulate data stored in Resource Description Framework (RDF) used by DBpedia

[6] LOCODE is a geographic coding scheme developed and maintained by United Nations Economic Commission for Europe.

When a new information referencing a port is acquired from a data source, the system needs to recognize which port this information concerns and assign an appropriate portid. This disambiguation is performed in the manner described in the following paragraphs.

Development of lexical resources. From the technical point of view, the disambiguation of port names, in its basic form, is very similar to the disambiguation of the flag names. Again, there is a lexical resource with pairs of port name variant portId. Having a collection of such pairs, stored in a form of database table, for any port name string extracted from a data source, the system searches through all pairs to find the matching port name variant and assigns the corresponding portId to the port string.

The lexical resource of the port name variants used in the SIMMO system, was obtained using the same approach as the one described for the flags, i.e. utilizing DBpedia lexicalization dataset. Using this procedure, for each port name it was possible to obtain its name variants. For example, for the port of Saint Petersburg in Russia the following port name variants were identified:

St Petersberg; St. Petersburg; Leningrad; Saint Peterburg; Sankt Peterburg; SanktPetersburg; St. Petersberg; Petrograd; SP; Saint-Petersburg; Saint Petersburg; St. Petersburg, Russia; Petersburg; St. Petersburgh; St.Petersburg; Piter; Sankt Petersburg; St Petersburg; St. Peterburg; Leningrad Saint Petersburg; Saint Petersburg, Russia; Saint Petersberg

Nevertheless, contrary to the flag name variants, in case of the ports some additional requirements had to be taken into account.

First of all, the names of ports are not unique. In many cases, there is more than one port with a given name (or a given name variant). For example, apart from St. Petersburg in Russia, there is a city with exactly the same name on Florida, USA, in which there is a port as well. As a result, if only the port name string is taken into account, it would be impossible to choose the correct port in other way than by chance. Sometimes, in a port name string there is an additional information about the country, in which the port is located (e.g. "St. Petersburg, Russia"). If correctly processed, this information may be used as an indication which port is the correct one. Still, if there is no such information, other approach must be used. In the next subsections the developed approaches for coping with this issue are presented.

Disambiguation of the home port based on vessel flag. At first, let's analyze a situation, in which a certain vessel in a data source has a home port name assigned and this name is not unique, e.g. Portsmouth (there are four ports with such name known to the SIMMO system). Thus, based solely on the port name string, the system doesn't know which port this information actually is referring to. To solve this issue, it was assumed that it is likely that the home port is located in the country associated with a flag of the analyzed vessel. Therefore, in such situations, the final portId is assigned according to the following procedure:

1 Get all portIds, for which a given port name string is a name variant,
2 Get list of countries, in which these ports are located, based on their LOCODEs (in the system each port has LOCODE assigned and two first letters of the LOCODE refer to a country),
3 Check if any of candidate ports (from the list from step 1) is located in the country associated with the flag of the vessel; if so, assign the corresponding portId as an identifier of the home port of the vessel being processed.

Using the described approach, we have processed the data extracted on ports extracted from internet sources. The described ambiguity was found in 2118 cases. Among them, in around 74,7% of cases the assigned flag of the ship was matching one of the ambiguous ports. This information was then used to decide which portId should be assigned. The flag of the country was not known in 14,3% of cases. In 10,95% of cases the flag associated with the given ship was not matching any of the possible home ports.

Disambiguation of the visited ports based on AIS messages and geographical proximity. Another scenario, when assigning the correct portId is challenging, is information about historical visits of vessels in ports. A list of vessel port calls with names of ports is retrieved from the internet sources (e.g. MarineTraffic). The port strings used in this data must be disambiguated a proper portIds must be assigned.

If it is unclear which port was actually visited by the vessel (e.g. name of the visited port is Portsmouth), information about geographical coordinates of the vessel at a given timestamp, taken from AIS messages, is used by to resolve the ambiguity. In this approach, a geographical proximity of the vessel to locations of different ports is calculated As a result, the port for which such proximity is the highest is selected and its portId is assigned to the data on visited ports.

Disambiguation of ports based on the port importance. In some cases, the approaches described in the previous subsections are not sufficient to correctly determine which port (out of those with a similar name) should be chosen during the disambiguation process. This may be for example due to the fact that there is no indication at all, which port is actually referenced to. For example, ports with the same name may be geographically very close to each other, as in the case of two Vancouver ports, just across the USA-Canadian border. In such case, proximity of a vessel to these ports may be insufficient to correctly determine, which port should be chosen during the disambiguation process.

For a human, in many situations it is obvious, after analyzing the available data, to which port the data is probably referring to. For example, Vancouver in Canada is a huge city and a very important port (47th largest container port according to World Shipping Council[7]), while a town with the same name in the United States is likely much less important, at least from the point of view of the maritime domain.

[7] http://www.worldshipping.org/about-the-industry/global-trade/top-50-world-container-ports, accessed 26 Jan 2016

Thus, similar reasoning was implemented in the SIMMO system. For this end, additional data about different ports (and cities associated with them) was utilized. Again DBpedia was used as a knowledge base, to which SPARQL queries about concepts (ports and cities) were sent in order to get values of DBpedia attributes, which potentially might be useful for determining the importance of the port and city. The obtained attributes included:

- populationTotal; population of the city, as a measure of the size of the city; it was assumed that usually ports in larger cities are of greater importance than for cities of smaller size,
- shipBuilder; the larger number of ships built in this city, the more important this port probably is,
- shipHomeport; if the port is a homeport for a larger number of vessels, it is probably more important as well.

Having these values, the system is able to choose the most important port based on the following heuristics. Each possible port is compared separately for these three values. Then the port which on average is on the highest position in the ranking is selected as the final disambiguated port.

Granularity of ports. Another difficulty with port disambiguation arises from the fact that ports may be perceived on different levels of granularity. Since the SIMMO uses only the main name of the city as the name of the port, there still may be smaller ports or docks in the area of the city, with names not containing the name of the main city but a name of a city district. For example, port name string "Hoogvliet" corresponds to a district of Rotterdam and in some data sources is provided as the name of the visited port. Still, the system's knowledge base there is only information about Rotterdam port and not about Hoogvliet. In such cases, portId of Rotterdam should be assigned to "Hoogvliet". However, often there are no mappings found in the DBpedia lexicalization dataset, which could be used in such scenario.

```
<geoname>
<toponymName>Gemeente Rotterdam</toponymName>
<name>Gemeente Rotterdam</name>
<lat>51.88246</lat>
<lng>4.28784</lng>
<geonameId>2747890</geonameId>
<countryCode>NL</countryCode>
<countryName>Netherlands</countryName>
<fcl>A</fcl>
<fcode>ADM2</fcode>
</geoname>
<geoname>
<toponymName>Hoogvliet</toponymName>
<name>Hoogvliet</name>
<lat>51.86333</lat>
<lng>4.3625</lng>
<geonameId>2753666</geonameId>
<countryCode>NL</countryCode>
<countryName>Netherlands</countryName>
<fcl>P</fcl>
<fcode>PPL</fcode>
</geoname>
```

Listing 2. An excerpt from the response of GeoNames Place Hierarchy for query about Hoogvliet place name

To resolve situation described above, GeoNames[8] web service is used. For each port name string, which the system was not able to disambiguate based on mappings from DBpedia lexicalization dataset, GeoNames Place Hierarchy web service is queried [9] in order to check, whether any of geographical units higher in the hierarchy to the given port can be found in the system's knowledge base (in the list of port names or port name variants) see the listing 2). If such geographical unit (port) is found, in the next step, it is checked whether the location of the analyzed unit is similar to the location of the known port identified in the previous step (as was previously mentioned, the location of known ports is stored in the system). If the coordinates are similar, it could be concluded that there is a suburb city relationship between the district (the processed port call) and the city associated with the port from the system's knowledge base. In such case, the name of the suburb is added as a name variant of the knowledge base, in the same way as it was done for mappings retrieved from DBpedia lexicalization dataset.

Evaluation of the port disambiguation process The proposed methods for the port name disambiguation were evaluated using the datasets extracted from the external data sources and stored in the SIMMO system. Below the results of the evaluation are presented.

Using the above-described approach, the system was able to assign portIds in 234710 cases out of 343610 records, in which a port string was specified. This is more than 68% of cases. The inability to assign portId to the remaining port strings may result from one of two reasons:

- an analysed port is not included in the system's knowledge base (e.g. a given port is a small port on a river) and thus disambiguation could not give any results,
- an analysed port is included in the system's knowledge base, but the disambiguation failed to identify it.

To check possible reasons, a sample of 150 port strings was randomly generated, out of all cases for which disambiguation failed. This sample was presented to human annotators which analysed, whether a port name string is present in the system's knowledge base. It turned out that only 12,8% of port strings without portIds assigned, were actually available in the knowledge base. Therefore, the developed methods failed to assign portIds in less than 5% of cases (12,8% out of 32% cases for which the portId was not assigned).

Further evaluation concerned checking to what degree the assigned portIds are correct. The evaluation of such disambiguation may be difficult, as in some situations there may be not enough data even for a human being to decide which port the given data actually corresponds to. Therefore, it was decided to run the evaluation only for the visits in ports. For this data, it was automatically checked what was the geographical distance between a given vessel and the port connected with a given portId at a defined timestamp. If the distance was relatively small then it may be assumed that the the correct portId was assigned.

[8] http://www.geonames.org/

[9] This we service returns all GeoNames higher up in the hierarchy of a place name. Source: http://www.geonames.org/export/place-hierarchy.html, accessed 12 Apr 2016

The figure 4 presents the accumulated distribution of the distances between position of the analysed vessel and the location of the disambiguated port. The median of the distances is 10.17 miles. In 90% of cases the distance was below 45 miles and in 95% of cases the distance was below 130 miles. For 150 miles, this value is settled on 97% and it does not rise with the further distance increase. While it is impossible to set a solid threshold to determine when the vessel actually is in a given port, the accuracy of port disambiguation using the defined methods may be evaluated as being between 90% up to 97%.

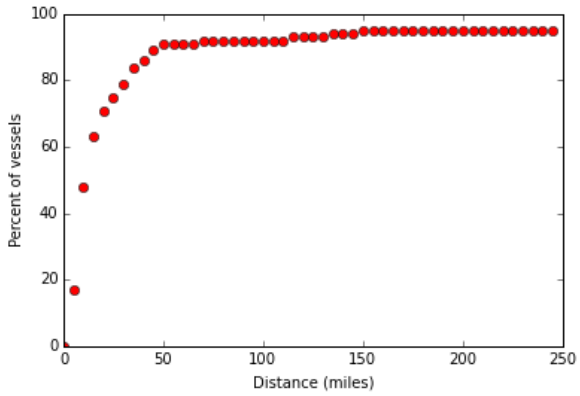

Figure 4. Distances between positions of the vessel and location of the disambiguated port which, according to the data, the vessel visited at the given timestamp

3.4 *Vessel types*

Each vessel may be described as being of a certain type, e.g. tug, fishing vessel, cargo etc. Usually, in a given data source, there is a predefined taxonomy of such vessel types, which is coherently used to describe vessels. Still, different data sources usually use different taxonomies. The same vessel type may be described using different strings, e.g. in a source A as "Fishing vessel", while in a source B simply as "fishing". Also, in different data sources vessel types may be perceived on different granularity levels or be a part of hierarchies that are somehow orthogonal to each other. This is a common problem when dealing with interoperability of different systems and data integration.

Again, if the data about vessel types is to be disambiguated automatically, simple relying on words used to describe such vessel types (strings) is not sufficient, and rather a unique identifier vesselTypeId should be assigned to each vessel type(for each word corresponding to any vessel type, regardless the source). Also, such identifier should be consistent across different data sources, so that even if in two data sources two different text strings are used to refer to the same vessel type, in the system the same vesselTypeId should be assigned.

A basic rule, which is used to determine whether two different strings from different sources refer to the same vessel type, is identification of vessel type name pairs used in data sources to refer to the same vessel. For example, let's assume that for a given vessel, in a source A its type is specified as "fishing" and in a source B as "fishing vessel". Let's also assume that the vessel type list, which is used in the system as the main list of the vessel types, is the one from the source A (further referred as the main list of vessel types). Finally, let's assume that it was identified that a certain number of ships which in the source A are assigned with "fishing" vessel type, in the source B are described as "fishing vessels". Thus, it may be expected that both strings refer to the same vessel type.

The reasoning described above was used in the SIMMO as a primary method for vessel types disambiguation. Still, using only one approach may be insufficient and some other methods should be used as well, e.g. string similarity measures between vessel type names.

```
['fishing vessel', 'fishing']: 1020 ['fishing', 'trawler']: 1156
['container ship', 'cargo hazard a (major)']: 1222
['passenger', 'ro-ro/passenger ship']: 1246
['crude oil tanker', 'tanker']: 1347 ['tanker', 'oil products tanker']: 1661 ['tanker', 'oil/chemical tanker']: 2086 ['cargo', 'container ship']: 2162
['cargo', 'general cargo']: 5536 ['cargo', 'bulk carrier']: 6309
```

Listing 3. The most frequent mappings between vessel types in two data sources used in the SIMMO system: Marinetraffic and Maritime Connector. The numbers to the right refer to the number of cases, when both vessel type names from two different data sources (values in brackets) refer to a vessel with the same shipId assigned.

What also have been taken into account is that in different sources the taxonomy of vessel type names may have a different granularity. For example, in the source A some vessel may be assigned a type "inland tanker", while in the source B, there is only a more general vessel type, "tanker". In such case, is-a relationship occurs, which is true only in one direction and false in the other. For example, it is true that each "inland tanker" is a "tanker", but it is false that each "tanker" is an "inland tanker". Therefore, a mapping is correct only if in the main list of vessel names, a more general vessel type name is specified. In such case, a string referring to more detailed vessel type can be used as a vessel type name variant for the more general type.

In order to identify such situations, the following heuristic was used. It is assumed that the longer name (i.e. consisting of a larger number of words) describes a more detailed entity. This assumption is based on an observation that additional words in vessel type name strings often restrict number of vessels that may be described using this name. For example, "oil/chemical tanker" is more detailed than a simple "tanker". Moreover to ensure that both vessel type names refer to similar concepts, it must be checked if the longer name contains the shorter one. For example, string "oil/chemical tanker" contains string "tanker". Thus, if, in a given mapping, it is identified that the more general term (the shorter string) is in the main list of vessel type names used in the system and the other vessel type name from the mapping contains this string, then this mapping may be used in disambiguation (the less general string may be used as a vessel type name variant for the more general one).

A list of the most common mappings, where the above-mentioned heuristics is used, is presented in Listing 4.

```
['tug', 'pusher tug']: 98
['tanker', 'lng tanker']: 101 ['tanker', 'bunkering tanker']: 119
['dredger', 'trailing suction hopper dredger']: 187 ['tanker', 'chemical tanker']: 222
['cargo', 'ro-ro cargo']: 342 ['tanker', 'inland tanker']: 397 ['tanker', 'lpg tanker']: 507 ['passenger', 'passengers ship']: 644 ['fishing', 'fishing vessel']: 1020
['passenger', 'ro-ro/passenger ship']: 1246 ['tanker', 'crude oil tanker']: 1347 ['tanker', 'oil products tanker']: 1661 ['tanker', 'oil/chemical tanker']: 2086 ['cargo', 'general cargo']: 5536
```

Listing 4. Vessel type mappings, filtered using a simple string similarity measure. The numbers to the right refer to the number of cases, when the two vessel type names in brackets referred in two different data sources to a vessel with the same shipId assigned.

Finally, manual analysis may be performed on other potential mappings by an expert and, based on that, additional mappings may be added to the system knowledge base.

3.5 *Classification societies*

Each vessel belongs to a classification society. The goal of the classification societies is "to provide classification and statutory services and assistance to the maritime industry and regulatory bodies as regards maritime safety and pollution prevention, based on the accumulation of maritime knowledge and technology"[10]. Names of classification societies, similarly to other data types, are expressed as strings and in each data source the same classification society may be referred to, using a different string. Therefore, for each acquired classification society name in the disambiguation process a proper identifier classId should be assigned. In the SIMMO system, there was an initial list of known classification societies with assigned classIds. This list was later extended during the disambiguation process.

```
['bureau veritas', 'nippon kaiji kyokai']: 22 ['american bureau of shipping', 'bureau veritas']: 29 ['det norske veritas', 'lloyds register']: 32 ['american bureau of shipping', 'lloyds register']: 41 ['dnv gl', 'germanischer lloyd']: 56
['registro italiano navale', 'american bureau of shipping']: 61
['korean shipping register', 'korean register']: 121
['dnv gl', 'det norske veritas']: 176
['lloyd\'s shipping register', 'lloyds register']: 267
['lloyds shipping register', 'lloyds register']: 605
```

Listing 5. The most frequent mappings between classification societies based on the fact that the same vessel was assigned different classification society strings in different sources. The results are much worse than for vessel types

The analysis of classification societies names started with generation of mappings in the same manner as it was done for flags and vessel types, i.e. by checking, if a singe vessel in different data sources has different classification society names assigned. However, in the case of the classification societies, this approach did not bring a lot of correct results, as shown on Listing 5; only a few of the most common mappings were correct and used in further analysis. This is probably due to the fact that vessels may change their classification society relatively often, in comparison to change of the vessel type (e.g. changing vessel type may require expensive modifications of the vessel itself). Therefore, different classification societies assigned to the same ship in different sources may result from the fact that information in one sources may be outdated in comparison to information provided in the other one.

Taking into account the obtained results, it has turned out that the number of distinct classification society names, for which the system was not able to assign classId based on the string comparison method, was only 192. Since, this number was relatively small, a manual analysis of the strings and assignment of the correct classIDs could have been performed. Based on the analysis, the system's knowledge base about the classification society name variants was updated. This allowed to disambiguate all classification society name strings.

3.6 *Company names*

In different data sources different strings may be used to refer to the same company. In many cases, such strings are similar, for example "Star Shipping Ltd" and "Star Shipping Limited". The aim of disambiguation in this case is to determine if two strings in fact refer to the same company and if so, assign the same identifier companyId to both of them.

In the first step, identification if different strings refer to the same company was performed by utilizing a string similarity measure, namely the Jaro distance [9]. Having two strings, this measure returns a numeric value between 0 and 1. The more similar the strings are, the higher value is returned.

The basic difficulty in the disambiguation of company names results from the fact that even for humans this task can be performed only with a limited certainty level (saying to what extend the output of the disambiguation is correct). It may be even more difficult to define how the term "single company" is understood and how to relate that to the analysis being performed. Let's analyze the following pair of company names: "Palmali Rostov, Russia" and "Palmali Shipping Services Instabul, Turkey". It is clear (at least for a human) that these strings refer to entities located in different countries. Still, after performing a search on the Internet, it may be learnt that both entities belong to the same group, Palmali Group of Companies[11]. In such case, classifying these two strings as the names of either the same company or two different companies depends on definition of a single company.

Still, in some cases, names of companies are not similar as far as Jaro measure is concerned, but still they may refer to the same company. For example, let's assume that we have the following strings: "U.S., Dept. of Transportation" and "USA Government - Washington DC, U.S.A". Jaro similarity between them

[10] See http://www.iacs.org.uk/document/public/explained/Class_ What-Why&How.PDF for details.

[11] http://palmali.com.tr/en/default.asp

is only around 0.54. Still, a human will notice that the Department of Transportation is a part of the USA Government. What is more, in the analysed data in 17 cases these two names were used in different data sources as the owners of the same ships.

The above mentioned example shows that string similarity measure in many situations is not sufficient to decide, whether two strings refer to the same company or not. Based on this observation, additional analysis was performed in which associations between the company names and the vessels was were identified. The analysis is similar to the one conducted for the vessel types. Again, for all ships it was analysed what company name strings are provided in different data sources for a given vessel (company name shipId company name mappings). If a certain pair of names occurs frequently in such analysis, it may be assumed that this pair refers to the same company. Still, similarly as in the case of the classification societies, the owner of a vessel may change relatively often, so if the data in different sources are outdated, the created mappings may be incorrect.

Taking all these aspects into account, it was analysed with what precision the automatic disambiguation of company names was performed. In the conducted experiment, different values of string similarity measure were set as a threshold for classifying two company names as referring to the same company. Two variants were analysed: 1) in which only the string similarity measure was used and 2) in which all pairs for which no ship was found, were discarded. Based on both variant, the company names can be either classified as referring to the same or to different companies.

To be able to evaluate the proposed approach, a sample of data was presented to human experts. The task was performed by three annotators. Each of them was presented with a collection of pairs of company names with different similarities between them. Also, for a part of these pairs, both strings were actually related to the same vessel, while for the other part not (the annotators did not know what was the similarity between strings and whether it was found in mappings or not). Each pair was annotated by exactly two annotators. To each pair, the annotators were to assign one of three values:

- both company names refer to the same company,
- company names refer to different companies,
- unknown (there is not enough information to decide which of the two other options should be chosen).

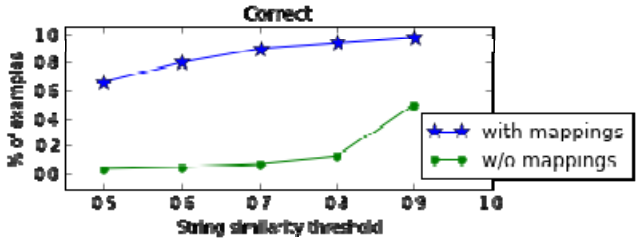

Figure 5. Precision of the proposed company disambiguation method for different thresholds on string similarity and for two variants: with or without additional filtering based on mappings found in the available data

Then exactly the same data sample was processed automatically by the system. The obtained results were compared with annotations produced by human experts to check the accuracy. Figure 5 presents results of the performed experiments. After setting a certain threshold for the string similarity, each pair of the company names, with similarity larger or equal to the threshold, may be classified as referring to the same company. Optionally, an additional filtering can be performed to discard all pairs which were not found in the database as referring to the same vessel.

The chart in Figure 5 presents what, according to the annotators, is the precision of classification[12]. Blue line presents the precision obtained for the pairs, for which at least one company name shipId company name mapping was found in the available data, while the green line corresponds to the pairs without this additional requirement. The chart clearly shows that the precision of the results obtained solely based on the string similarity is very low. Even after setting a very high threshold, it is not higher than 0.5. Utilization of company name shipId company name mappings allows to dramatically increase the precision, even for much lower thresholds.

Based on the experiments, the disambiguation of company names with threshold equal to 0.7 was conducted. As a result, only for pairs found in the identified mappings, the precision of disambiguation accounted to the level of 90%. Using this approach, it was possible to assign IDs to 11525 out of 115419 records, what constitutes around 10% of company names found in the data.

4 SUMMARY AND FUTURE WORK

The process of disambiguation of named entities is the basic task, which need to be performed to integrate data coming from heterogeneous internet data sources and to enable further analysis of the integrated data. In the article various approaches to disambiguation of the named entities related to maritime domain are presented. Using the developed approaches, for some types of entities, the disambiguation could have been performed with the high accuracy. It concerns inter alia ports, flags, vessels and classification societies.

Still, for the other types of entities, like maritime-related companies or vessel types, there is a need for a further research and development of methods, which would provide a more precise fusion of data. For the vessel types, probably a different data model (e.g. a taxonomy with is-a relationships) could be used. However, it would require a more prolonged engagement of the domain experts. For the disambiguation of the company names, an additional reasoning may be implemented, which would utilize data from additional sources, being an indication of what strings are used to reference the same company. Still, according to the performed evaluation, it may be concluded that in general the presented approaches

[12] Precision is understood as a ratio of pairs correctly classified by the system as referring to the same company to all pairs classified as such

may be successfully utilized in similar systems in the future.

ACKNOWLEDGEMENT

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

International Maritime Organisation: The International Aeronautical and Maritime Search and Rescue (IAMSAR) Manual. IMO/ICAO, London (2013)

el Pozo, F., Dymock, A., Feldt, L., Hebrard, P., di Monteforte, F.S.: Maritime surveillance in support of csdp. Technical report, European Defence Agency (2010)

Angerman, W.S.: Coming full circle with boyd's ooda loop ideas: An analysis of innovation diffusion and evolution. Technical report, DTIC Document (2004)

Vassiliadis, P.: A survey of extract–transform–load technology. International Journal of Data Warehousing and Mining (IJDWM) 5(3) (2009) 1–27

Abramowicz, W., Eiden, G., Małyszko, J., Stróżyna, M., We˛cel, K.: SIMMO Project. Deliverable 1.2 Report on selected internet data sources, defined cooperation models and intelligence analysis scenarios. Research report, Poznan´ University of Economics, LuxSpace Sarl (2015)

Bilenko, M., Mooney, R.J.: Adaptive duplicate detection using learnable string similarity measures. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '03, New York, NY, USA, ACM (2003) 39–48

Rahm, E., Do, H.H.: Data cleaning: Problems and current approaches. IEEE Data Eng. Bull.
23(4) (2000) 3–13

Alberga, C.N.: String similarity and misspellings. Commun. ACM 10(5) (May 1967) 302– 313

Jaro, M.A.: Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. Journal of the American Statistical Association 84(406) (1989) 414–420

Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate record detection: A survey. Knowledge and Data Engineering, IEEE Transactions on 19(1) (2007) 1–16

Wentland, W., Knopp, J., Silberer, C., Hartung, M.: Building a multilingual lexical resource for named entity disambiguation, translation and transliteration. In: LREC. (2008)

Vespe, M., Sciotti, M., Battistello, G.: Multi-sensor autonomous tracking for maritime surveillance. In: Radar, 2008 International Conference on, IEEE (2008) 525–530

Kazemi, S., Abghari, S., Lavesson, N., Johnson, H., Ryman, P.: Open data for anomaly detection in maritime surveillance. Expert Syst. Appl. 40(14) (2013) 5719–5729

Kaczmarek, T., Węckowski, D. 347. In: Harvesting Deep Web Data through Produser Involvement. IGI Global (2013) 200–221

Chang, K.C.C., He, B., Li, C., Patel, M., Zhang, Z.: Structured databases on the web: Observations and implications. ACM SIGMOD Record 33(3) (2004) 61–70

Rhodes, B.J., Bomberger, N.A., Seibert, M., Waxman, A.M.: Maritime situation monitoring and awareness using learning mechanisms. In: Military Communications Conference, 2005. MILCOM 2005. IEEE, IEEE (2005) 646–652

Helldin, T., Riveiro, M.: Explanation methods for bayesian networks: review and application to a maritime scenario. In: Proc. of the 3rd Annual Skövde Workshop on Information Fusion Topics (SWIFT 2009). (2009) 11–16

Mano, J.P., Georgé, J.P., Gleizes, M.P.: Adaptive multi-agent system for multi-sensor maritime surveillance. In: Advances in Practical Applications of Agents and Multiagent Systems. Springer (2010) 285–290

Ding, Z., Kannappan, G., Benameur, K., Kirubarajan, T., Farooq, M.: Wide area integrated maritime surveillance: An updated architecture with data fusion. In: Proceedings of the Sixth International Conference of Information Fusion, Australia. Volume 2. (2003) 1324–1333