Artificial intelligence for ocean science data integration

current state, gaps, and way forward

Sagi, Tomer; Lehahn, Yoav; Bar, Koby

Published in:
Elementa: Science of the Anthropocene

DOI (link to publication from Publisher):
10.1525/elementa.418

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

ELEMENTA
Science of the Anthropocene

Sagi, T, et al. 2020. Artificial intelligence for ocean science data integration: current state, gaps, and way forward. *Elem Sci Anth*, 8: 21. DOI: https://doi.org/10.1525/elementa.418

PRACTICE BRIDGE

Artificial intelligence for ocean science data integration: current state, gaps, and way forward

Tomer Sagi[*,†], Yoav Lehahn[‡] and Koby Bar[*]

Oceanographic research is a multidisciplinary endeavor that involves the acquisition of an increasing amount of in-situ and remotely sensed data. A large and growing number of studies and data repositories are now available on-line. However, manually integrating different datasets is a tedious and grueling process leading to a rising need for automated integration tools. A key challenge in oceanographic data integration is to map between data sources that have no common schema and that were collected, processed, and analyzed using different methodologies. Concurrently, artificial agents are becoming increasingly adept at extracting knowledge from text and using domain ontologies to integrate and align data. Here, we deconstruct the process of ocean science data integration, providing a detailed description of its three phases: discover, merge, and evaluate/correct. In addition, we identify the key missing tools and underutilized information sources currently limiting the automation of the integration process. The efforts to address these limitations should focus on (i) development of artificial intelligence-based tools for assisting ocean scientists in aligning their schema with existing ontologies when organizing their measurements in datasets; (ii) extension and refinement of conceptual coverage of – and conceptual alignment between – existing ontologies, to better fit the diverse and multidisciplinary nature of ocean science; (iii) creation of ocean-science-specific *entity resolution* benchmarks to accelerate the development of tools utilizing ocean science terminology and nomenclature; (iv) creation of ocean-science-specific schema matching and mapping benchmarks to accelerate the development of matching and mapping tools utilizing semantics encoded in existing vocabularies and ontologies; (v) annotation of datasets, and development of tools and benchmarks for the extraction and categorization of data quality and preprocessing descriptions from scientific text; and (vi) creation of large-scale word embeddings trained upon ocean science literature to accelerate the development of information extraction and matching tools based on artificial intelligence.

Keywords: Data Integration; Artificial Intelligence; Ontologies; Oceanography

1. Introduction

The study of the ocean is one of the biggest scientific challenges of the 21st century. It has a direct impact on our understanding of Earth's climate (Stocker et al., 2013) and biogeochemical cycling (Field et al., 1998), as well as on our ability to provide human society with food, chemicals, and energy (Lehahn et al., 2016). Oceanographic research strongly relies on in-situ and remotely-sensed observations, which describe physical, chemical, and biological seawater properties at a given time and place. These observations are collected from various crewed and autonomous platforms, including research vessels, floats (Roemmich et al., 2009), drifters (Lumpkin et al., 2017), autonomous vehicles (Eriksen et al., 2001), and satellites (Lehahn et al., 2018), providing an abundance of interdisciplinary information on processes occurring over a wide range of spatial (from micrometers to thousands of kilometers) and temporal (from seconds to decades) scales.

Over the last century, numerous in-situ and remotely-sensed measurements have been taken, resulting in the creation of an increasingly large amount of oceanic data. In recent years, with the enhanced utilization of satellites and autonomous observation platforms, these data are collected at a blistering rate. Improving the scientific community's ability to integrate, share, and explore this vast amount of data is an urgent task that will contribute substantially to our understanding of the ocean and its role in the Earth system.

Several public data repositories have emerged to enable the archiving and sharing of data collected between researchers. For example, PANGEA (2020), a data repository for publishing and distributing georeferenced data

* Department of Information Systems, University of Haifa, Haifa, IL
† Department of Computer Science, Aalborg University, Aalborg, DK
‡ Department of Marine Geosciences, Charney School of Marine Sciences, University of Haifa, Haifa, IL
Corresponding authors: Tomer Sagi (tsagi@is.haifa.ac.il); Yoav Lehahn (ylehahn@univ.haifa.ac.il)

from Earth system research, hosts more than 370,000 datasets. The National Centers for Environmental Information (National Oceanic and Atmospheric Administration, 2020b) stores over 25 petabytes of atmospheric, coastal, oceanic, and geophysical data. Copernicus (European Commission, 2020) archives datasets from several domains such as marine, climate, and agriculture, as part of a European Union program for observing the Earth. The extensive availability of data repositories provides oceanographic researchers with the ability to tap into a multitude of data collected by their peers and use it in their own studies.

One of the main obstacles for a researcher when compiling data from existing data sources is to overcome the semantic distance between datasets. Thus, when conducting such research, there is a need for manual data integration work done by an expert. In a recent review (Gregory et al., 2019), the authors described some of the challenges facing researchers when manually integrating data from multiple disparate studies.

Data integration is the art and science of reconciling two or more collections of data with each other. Data integration is as old as data. In 1975, the National Bureau of Standards and the Association for Computing Machinery issued the recommendation that, when integrating data from digital and physical files into the newly standardized Database Management Systems (DBMS), practitioners should maintain a data-dictionary to enable efficient and effective data integration (Berg, 1976). With the emergence of the federated database (Hammer and McLeod, 1979), a database composed of multiple independent database systems, the need for a central mediated schema was created. A secondary problem created by federated databases was the prevalence of unwanted data duplication between the systems. The advent, and subsequent popularity of the World Wide Web, brought about a host of new opportunities for sharing data, providing portals and services based on the integration of data from multiple sources covering the same domain, such as the domain of travel reservation (e.g., www.orbitz.com). The process of data integration began as a manual one (Goodhue et al., 1992), gradually transitioning to a semi-automated process supported by software tools. The arrival of Big Data has increased both the number and sizes of available data-sources, bringing about additional challenges and opportunities for data integration (Dong and Srivastava, 2015).

We are at a time where artificial intelligence (AI) is applied ubiquitously across scientific domains and disciplines. First and foremost of AI research fields is the field of machine learning (ML), the science of building software that learns from experience. Recent years have seen a concurrent increase in data (serving as experience for ML) and available cloud computing solutions to utilize the data. These phenomena, together with the arrival of deep learning (DL) as an efficient and effective method for ML, have caused ML to expand into an increasing number of fields (Jordan and Mitchell, 2015). Pioneered by Doan et al. (2002), the use of ML in data integration has been expected for some time now (Halevy et al., 2006). Recently, widespread use of ML in data integration appears to be the new norm (see review by Dong and Rekatsinas, 2018). Concretely, ML has been used to create weighted ensembles of schema matchers (Gal and Sagi, 2010), map relational databases into ontologies (De uña et al., 2018), and create sub-groups of records to speed up entity resolution (see review by O'Hare et al., 2019). However, the advances in data integration and specifically AI-assisted data integration have been utilized sparingly, if at all, in the ocean sciences.

In this paper, we systematically deconstruct the process of integrating a multitude of datasets in the ocean science domain into specific phases and tasks. For each task, we review state of the art in AI-assisted data integration and discuss the barriers and challenges to its adoption in the ocean sciences. We begin in the following section by formally defining and providing background on artificial intelligence, data integration, and how they are used together. We then present our model of data integration processes in ocean science and how artificial intelligence can support these efforts. To demonstrate the implications of having ocean-science-specific-AI tools, we then describe and provide results from an automated entity extraction task on oceanic datasets.

2. Background and definitions

Before we dive into the use of artificial intelligence for data integration in ocean sciences, we review data integration (DI), artificial intelligence (AI), and the use of AI techniques in DI.

2.1. Data integration

DI is the process of combining two or more datasets. Datasets are collections of structured data described by a *data description*, also known as a *schema*. A dataset may be simple as a table, with rows as data and the header row as a schema, or complex as a netCDF (UNIDATA, 2019) file containing numerical matrices.

Figure 1 reviews the five components of the DI process. *Schema matching* (1) aligns two or more schemas to find correspondences between them (see survey by Shvaiko and Euzenat, 2013 and books: Gal, 2011; Bellahsene et al., 2011). *Schema mapping* (2) operationalizes these correspondences into data-transformation functions (e.g., Alexe et al., 2011). *Entity resolution* (3) is the task of identifying different instances related to the same entity (see surveys Papadakis et al., 2016; O'Hare et al., 2019). *Entity consolidation* (4) is the process of merging all data about the same entity coherently (e.g., Hogan et al., 2012). An orthogonal but crucial component of the DI process is *data cleansing* (5), which can be applied to both the original data and the merged dataset (Abedjan et al., 2016).

Note that entity consolidation is designed for database records, where each property has a single value. Most oceanic datasets are comprised of both database-style records recording a dataset's metadata and a large series of numbers varying over geographical or temporal dimensions. Integrating the numerical component, introduces two new dimensions to the integration process, namely resolution and distance. Numerical analysis and model building requires a continuous set of numbers with the same resolution. For example, satellite images might have a spatial resolution of 1 km and a temporal resolution of one day, while a buoy in the same area and time has a

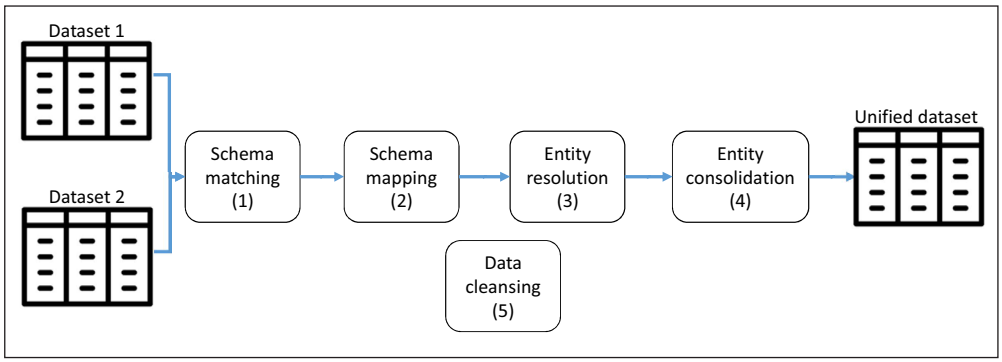

Figure 1: The process of data integration. The data integration process takes two datasets and combines them into a unified dataset by performing five composable tasks. Schema matching (1) aligns the schemas of the two datasets. Schema mapping (2) performs any transformations required by the different semantic of the aligned fields. Entity resolution (3) identifies duplicate records, and entity consolidation (4) merges them. Data cleansing (5) can be applied at any point to detect and correct errors. DOI: https://doi.org/10.1525/elementa.418.f1

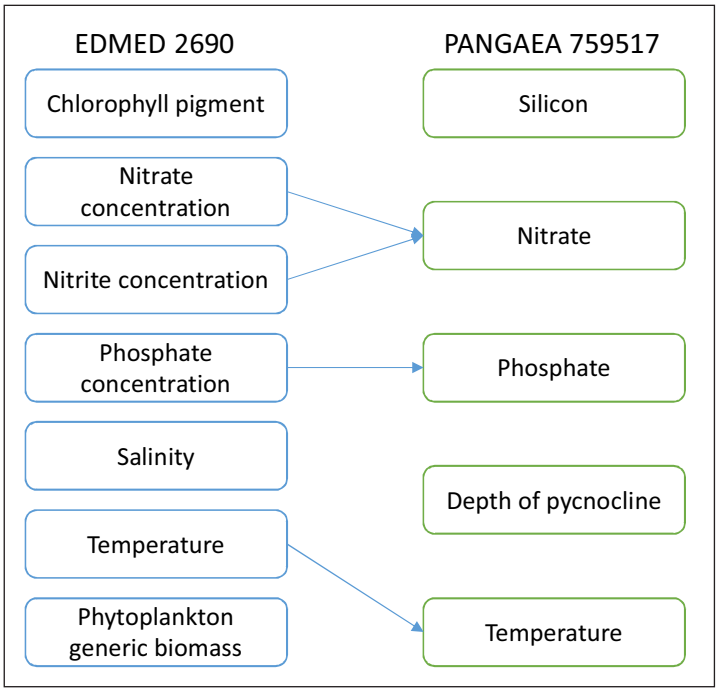

Figure 2: Schema matching of two oceanic datasets. The figure shows correspondences created by a schema matching process between the schema of an EDMED dataset and that of a PANGAEA dataset. DOI: https://doi.org/10.1525/elementa.418.f2

pinpoint spatial resolution but may often lay a few kilometers away from the nearest sea surface image edge, due to cloud cover. To build an integrated model over both sets, one must perform interpolation and extrapolation and assess the reliability of their model given these differences and the methods employed to bridge them. Multi-sensor data fusion techniques (Lesiv et al., 2016; Waltz and Waltz, 2017) have diversified and grown from statistically based methods to more elaborate ML-based methods. In the interest of brevity and focus, we limit the exploration of this task in the rest of this paper, leaving it for future work.

Example 1 ***Schema matching and mapping (Figure 2).*** *A researcher wishes to integrate PANGAEA dataset 759517 (semina and Mikaelyan, 1994) with dataset 2690 stored on EDMED (British Oceanographic Data Centre, 2020).* ***Figure 2*** *presents the correspondences between the two datasets' schemas, a result of a manual schema matching process. A note added to the Nitrate field of the PANGAEA dataset identifies this field as actually measuring the sum of nitrates and nitrites, justifying the correspondences to the Nitrite and Nitrate fields in the EDMED dataset. This double correspondence can be converted later to a sum of the two values in the schema mapping process to convert data points from these fields under the EDMED schema to the PANGAEA schema.*

Example 2 ***Entity resolution.*** *Consider* ***Table 1*** *where the same data point is presented from the diatom data integration effort by Leblanc et al. (2012) (first row) and one of its constituent datasets, a supplement to Assmy et al. (2007) (second row). We manually schema-matched and mapped the second row to the first row's schema; however, it is still unclear if indeed, these represent the same data point. For*

Table 1: Entity resolution: two records mapped into the same schema. DOI: https://doi.org/10.1525/elementa.418.t1

Project ID	Cruise or station ID	Date	Longitude	Latitude	Name entry
EISENEX	out of +Fe patch *st°*108	11-29-2000	20.60	–47.67	*Thalassionema nitzschoides <20 μm*
European iron enrichment experiment in the Southern Ocean (EisenEx)	PS58/108-1 (CTD149)	2000-11-29T15:33:00	20.64733	–47.66817	*Thalassionema nitzschioides* var. *lanceolata*, biomass as carbon *[μg/l]* (*T. nitzschioides* var. *lanceolata C*)

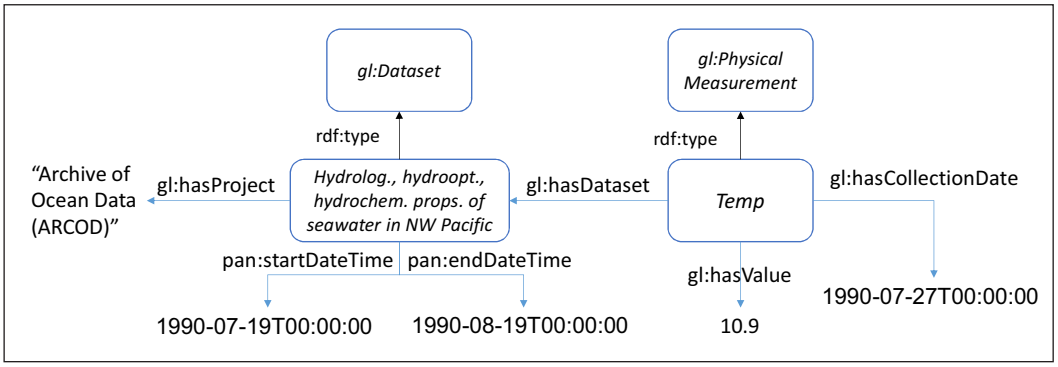

Figure 3: An example knowledge graph. In the figure, a graph fragment with some of the data from Semina and Mikaelyan (1994) is presented in machine-readable manner by using well-defined ontological and schematic properties that have well-defined relations to other properties. These definitions and properties allow integrating these data with data from other datasets. Boxes represent entities, quoted strings are literals, and edges represent predicates that connect a subject (entity) to an object (entity/literal). Prefixes denote the ontology/schema in which the property/class are defined, with rdf denoting the resource description framework (RDF) schema (https://www.w3.org/TR/rdf-schema/), gl denoting the geolink ontology http://schema.geolink.org/), and pan denoting the PANGAEA schema. The entity *Temp* represents a data point and is connected to its parent dataset via a gl:hasDataset predicate. The data point is connected to the collection time via a gl:hasCollectionDate predicate, and the dataset is connected to its temporal coverage through the predicates pan:startDateTime and pan:endDateTime. Both entities (i.e., data point and dataset) are connected to their ontological classes via an rdf:type predicate. The dataset entity is connected to a literal describing its project. DOI: https://doi.org/10.1525/elementa.418.f3

large datasets, the entity resolution task may be daunting, requiring n^2 comparisons where n is the number of records over all datasets. Thus, common approaches perform a process of blocking, where records are grouped by (one or more) shared properties. In our example, these two data points were part of a dataset containing 293,000 data points, of which more than half may be duplicates. To avoid performing 8.6×10^{11} *comparisons, we could first group records by the longitude, latitude, depth, and date, and then perform comparisons only within each group (block in entity resolution terms).*

Entity resolution can occur at different levels of granularity and for different entities appearing in the dataset. The example given above identified the same data item in the two datasets, similarly, the authors were required to resolve different diatom species described differently. In the authors' own words: "In total, 1364 different taxonomic entries were found, but were reduced to 727 different taxonomic lines...."

2.2. Artificial intelligence

Kaplan and Haenlein (2019) define AI as: "a system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation". The definition encompasses three core aspects of AI systems. *Interpretation* of external data requires reasoning, i.e., deriving conclusions from raw inputs using an internal representation of knowledge. *Learning* from data is the ability to change a system's internal model based upon examples. *Adaptation* means the system can perform actions that change according to a change in the internal representation. In the following we describe the first two core aspects and their supporting technologies. The third aspect targets autonomous agents, such as robots, and game-playing (e.g., Silver et al., 2016), which are not relevant to the task of data integration and therefore are not addressed further.

2.2.1. Knowledge representation and reasoning systems

Allowing computer software to reason requires a way to represent and store large amounts of knowledge, and systems able to query knowledge and reason over it. One of the most mature approaches, backed by substantial commercial and academic effort, is that of the Semantic Web as envisioned by Berners-Lee and Hendler (2001). Under this conceptual model, knowledge graphs (KG) have become a standard for representing facts. As their name suggests, KG are a network-based representation, where entities and literals are nodes, and predicates or relations are the edges.

Example 3 *In* ***Figure 3****, a knowledge graph fragment presents our knowledge about a data point from a dataset*

(Semina and Mikaelyan, 1994) stored on PANGAEA. The dataset entity (Hydrolog...) is connected via the predicate gl:hasProject to a literal describing it. The data point entity (Temp) is connected via a predicate gl:hasDataset to the dataset entity describing the fact that the former is a component of the latter.

In general-purpose knowledge graphs such as Wikipedia-based DBpedia (Auer et al., 2007), entities may represent people, places, and abstract things, such as events, while literals represent single pieces of information such as names, titles, and dates. Ontologies provide a conceptualization of the domain (or domains) described by the knowledge graph, adding entailment mechanisms such as the ability to group entities into a class, create *same-as* links between entities, equivalence relationships between classes, and denote predicates as sub-properties. For example, both entities in the example above are connected via *rdf:type* predicates to their ontological classes. These two entities and the predicates prefixed with *gl:* are defined in the GeoLink base ontology (Krisnadhi et al., 2015). The definition of an *rdf:type* is specified in the resource description framework (RDF) and can be found at https://www.w3.org/TR/rdf-schema/. Querying information represented as a KG is often done using SPARQL (Prud'hommeaux and Seaborne, 2008), a data retrieval language enhanced with semantic inference constructs.

2.2.2. Machine learning

Endowing software with the ability to learn from examples has been studied extensively over the past 60 years. ML has been used to automate tasks over the entire expanse of the human endeavor from predicting relations in knowledge graphs (see review by Nickel et al., 2016) to forecasting solar radiation (Voyant et al., 2017). Machine learning techniques can be broadly divided into two types, *supervised* and *unsupervised* by the type of input provided to the learning algorithm.

Unsupervised learning techniques provide the learning algorithm with a large collection of items sampled from the target population and some target metrics to assess the quality of the task result, leaving the algorithm to attempt and optimize these quality criteria. Classic examples include clustering techniques such as K-Means (Hartigan and Wong, 1979). The effectiveness and applicability of using unsupervised techniques to learn a representation have increased significantly with the appearance of large amounts of user-generated content on the Internet. For more details, see the seminal paper on the unreasonable effectiveness of data by Halevy et al., (2009). A similar opportunity exists in oceanic sciences with the increasing availability of large amounts of autonomously collected and remotely sensed data (see Durden et al., 2017 for a review).

Supervised learning techniques require a (hopefully large) set of tagged examples. For example, to identify the semantic information conveyed by a set of numbers representing the pixels in a picture, a supervised ML algorithm requires a set of pictures labeled as *cats*, another labeled as *dogs*, etc (Russakovsky et al., 2015). Similarly, to identify people and places mentioned in a text, an ML model requires sentences where they are clearly labeled as such. Given a metric to which the ML's prediction can be compared to the real tag, the ML algorithm can alter its internal representation to achieve better results on the task at hand. For example, using a quadratic loss metric, calculated over the distance between the final result vector and the expected one, is common in computer vision tasks. However, obtaining tagged examples is often difficult and expensive, as it requires humans, often experts, to tag the examples. Furthermore, one needs to obtain a set of examples which is representative of the target task. More often than not, the examples on which ML-models are trained are those for which gathering information is more convenient than representative.

2.2.3. Information extraction

The ability of AI systems to obtain information from raw data relies upon three fields of research. *Computer vision* (e.g., Krizhevsky et al., 2017) aims to extract meaning from images and video, *(textual) information extraction* focuses on text (e.g., Martinez-Rodriguez et al., 2020), and *audio (speech) recognition* (e.g., Hinton et al., 2012) converts sound into more meaningful information such as text and emotion markers (Schmidt and Kim, 2011).

2.3. AI in data integration

2.3.1. Ontology-based data integration and access

Taking advantage of the AI knowledge representation and inference mechanisms, *ontology-based data integration* (OBDI) uses ontologies to consolidate several heterogeneous sources into one source (see review by Ekaputra et al., 2017). For example, if the schema in one dataset contains the specific instrument (e.g., CTD/Rosette) and in another the instrument type (e.g., Cast), we could use the *hasType* ontological construct to integrate them.

In many cases existing data sources are not linked to an ontology, rendering OBDI impossible. *Ontology-based data access* (OBDA) is an alternative model that provides access to the data layer through a declarative mapping between autonomous data layers and a domain-specified ontology (Xiao et al., 2018). A typical development process of an OBDA system for a project that has a standard, non-ontological database will contain the following steps. (a) Create an ontology of domain-specific user knowledge. (b) Write mapping that connects (usually through SQL queries) the ontology to the project's database. (c) Write a query using ontology's vocabulary as a semantic query language query, such as SPARQL. (d) Build an OBDA system framework that automatically rewrites the SPARQL query to the query language in which the project's database operates.

2.3.2. Word embeddings

Early work in DI heavily relied upon measures such as Jaccard similarity (e.g., He and Chang, 2006) and n-gram techniques (e.g., Do and Rahm, 2002) to ascertain if two strings are similar. However, syntactic methods ignore the semantics, or meaning, of words. Such techniques can find *plane* and *airplane* to be similar, but not

plane and *aircraft*. To overcome this weakness, thesauruses such as WordNet, and later Wikipedia, were introduced. However, these techniques required accurate spelling and were often baffled by technical terms and abbreviations.

The appearance of *word embeddings* has revolutionized the approach towards word, phrase, and sentence similarity. Word embedding was originally designed to convert text to the numerical representation required by DL techniques. The technique represents each word in the vocabulary with a d-dimensional vector of real numbers $w \in \mathbb{R}^d$. Word embedding has been extensively used in AI applications as an underlying input representation that serves as a word dictionary and enables better capture of the semantic meaning of the word (Levy et al., 2015). The following hypotheses have been noted (Bolukbasi et al., 2016). (a) Vectors of words of similar meaning tend to be closer. (b) The vector differences between vectors representing word embeddings have been shown to represent relationships between words. A famous example is the male/female relationship captured by the word2vec implementation of word embedding, where Mikolov et al. (2013) showed that $\overrightarrow{King} - \overrightarrow{Man} + \overrightarrow{Woman} \approx \overrightarrow{Queen}$.

Thus, a word would be embedded in a high-dimensional space as a vector, and a sentence became a collection of such vectors. Word similarity now becomes a problem of vector similarity. Useful embeddings are those that place similar words close to each other in this high-dimensional space. Embeddings are learned from large collections of text, in an unsupervised manner. Thus, they can be fine-tuned to a specific domain by retraining some of the embeddings on a collection of domain-representative documents. Popularized by Word2Vec (Mikolov et al., 2013), recent methods include GloVe (Pennington et al., 2014), Flair (Akbik et al., 2018), and BERT (Devlin et al., 2019). The latter two use character-based embedding, which can also overcome spelling and abbreviation issues.

2.3.3 Machine learning for data integration

The use of machine learning for schema matching had been pioneered by Doan et al. (2000), followed by work by Gal and Sagi (2010). In both cases, machine learning was used to learn an ensemble model or method to combine the results of multiple matchers by training the ensemble method on the results of previous matching attempts. Sagi and Gal (2013) took this method one step further by learning to adapt the ensemble weights according to the results of the actual matching performed at run-time. Thus, the features upon which their model was trained were not the choice of matchers, but rather the structure and various counting statistics of the match result. Recently, word embeddings were used to enhance the effectiveness of schema matchers by Fernandez et al. (2018).

ML techniques have been used for entity resolution as well. Kenig and Gal (2013) used an unsupervised ML technique called *maximal frequent item-sets* (MFI) to learn the optimal clusters in which to search for duplicates. Sagi et al. (2017) expanded upon this work by training an alternating decision tree model (Freund and Mason, 1999) to classify pairs within the blocks to matched and unmatched entities. Recent work, such as by Ebraheem et al. (2018), utilizes word embedding to create semantically similar clusters as well as recommend matched pairs. Data tamer (Gubanov et al., 2014) uses ML for entity consolidation by predicting which data item is most likely to be relevant.

3. Data integration in ocean science

In this section, we formalize the data integration process for oceanic datasets. Under this formalization, we can compare similar tasks and examine tools employed in support (or in relief) of the extensive manual labor otherwise required. After describing each step, we review current work in ocean science and list the remaining gaps accompanied by specific directions for future work.

A data integration project can be described as having three major phases (**Figure 4**, top layer). In the *Discovery* phase, the list of possible candidate datasets for the project is compiled. In the *Merge* phase, candidate datasets are harmonized semantically, computationally, and geographically to form one large and coherent dataset. In the *Evaluate/Correct* phase, the results are analyzed to assess quality, coverage, and bias, and appropriate corrections are made to support assertions made over the data.

In the following sections, we describe these phases in detail, further dividing them into distinct steps. Although the integration process described holds whether done manually or automated, we point out how the DI technologies described in Section 2 can be used to automate the different steps, allowing to scale such projects and integrate large amounts of data. Where appropriate, we describe how AI technologies can in-turn support the DI processes. The bottom two layers of **Figure 4** summarize these supporting relationships.

3.1. Discover

Data discovery is the phase where candidate datasets are collected to fit a set of study parameters. For example, Luo et al. (2012) searched for datasets containing sampling of marine N_2 (dinitrogen) fixing organisms. Similarly, Wang et al. (2017) focused their efforts on geochemical data. The process of data discovery can be divided into three distinct steps, namely, *search*, *link*, and *identify*, described below.

3.1.1. Search

In the search step, a list of candidate research is collected. Search is performed on repositories or through portals that provide access to multiple repositories, hereafter referred to as *data sources*. Data sources may contain either textual descriptions of studies (i.e., scientific papers) or the datasets themselves. Google Scholar is an example of a scientific portal to study descriptions, while PANGAEA is a repository of datasets.

When searching for relevant research, users use search tools provided by the data sources. These tools can be classified into one of three types of interfaces. *Key word* queries comprise a sequence of terms of which at least one should be present in the dataset for it to be returned in the results. *Ontological* queries rely on well-defined

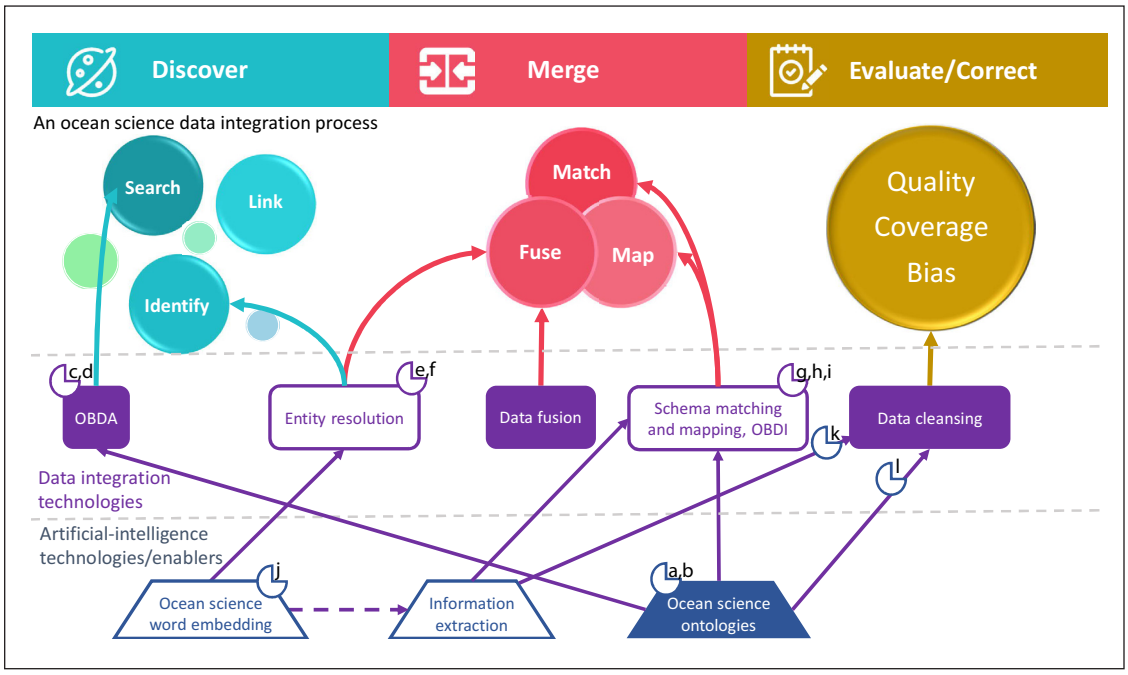

Figure 4: The three phases of the data integration process, and their application in ocean science. The top layer describes the process: in the *discover* phase, a list of candidate datasets with possible relevancy to the project is compiled; in the *merge* phase, candidate datasets are harmonized semantically, computationally, and geographically to form one large and coherent dataset; in the *evaluate/correct* phase, an analysis of the resulting dataset is performed to assess quality, coverage and bias, followed by appropriate corrections that are made to support assertions made over the data. The middle layer shows how data integration technologies support the process. OBDA and OBD stand for *ontology-based data access* (A) and *integration* (I) respectively. The bottom layer contains three AI technologies/enablers that support the data integration technologies. Full-colored rectangles and trapezoids represent technologies/enablers in current use. Outline-colored-only shapes represent technologies and enablers that are not currently in use in ocean science data integration. Additional gaps are listed as lower-case letters corresponding to the gaps listed in Table 3. DOI: https://doi.org/10.1525/elementa.418.f4

ontological terms such as organism species or molecular compounds, which the user specifies together with logical constraints and entailment allowances to form a logical statement. Each candidate result must satisfy the logical statement to be returned. *Parameter* queries rely on metadata associated with the research, such as the publication date or the geographical location of the samples collected. Queries are formed by defining restrictions and combining them using simple logical operators (and/or/not). To exemplify the difference between ontological search and parameter search, consider the following.

Example 4 *A researcher is interested in datasets containing measurements of phytoplankton biomass, among other parameters. In a parameter search, that researcher would be required to search for all possible subgroups and types of phytoplankton, such as diatoms, Fragillariophyceae, and Coscinodiscophyceae, and then collate the results. In an ontological search, the researcher can simply ask for all diatoms and specify that they wish for all sub-species as well, then receive all datasets containing the biomass of a species present in the taxonomic tree under diatoms. However, to support such a search, each parameter defined over a dataset needs to be aligned correctly with a comprehensive ontology, a task that is daunting when done retrospectively over large collections of datasets.*

Table 2 provides a partial list of data sources, oceanic research portals and repositories current to January 2020, their type (R stands for Repository and P for Portal), and the extent to which they support the search tools described above (all data sources listed provide key-word search). A notable omission from this list is the set of commercial cloud services participating in NOAA's Big Data Project (National Oceanic and Atmospheric Administration, 2020a). Access to this data source is rudimentary, and the number of datasets provided is limited.

Taxonomies are widely used in the ocean sciences (Claramunt et al., 2017). Some examples are World Register of Marine Species (WoRMS Editorial Board, 2020) that holds a detailed taxonomy of marine species, AlgaeBase (Guiry and Guiry, 2020), a global algal database, and FishBase (Froese and Pauly, 2020). An *ontology* is an explicit specification of a conceptualization that defines the terms in the domain and relations among them (Gruber, 1995).

All ontologies use some form of *vocabularies* in order to express terms and specify their meanings (Uschold, 1998). Similarly to taxonomies, they adopt a classification structure. However, ontologies add properties for each class and a set of axioms and rules that allow reasoning and full domain conceptualization (Zeng, 2008). Leadbetter et al. (2010) provide a systematic review of ontologies for the maritime domain. A few notable mentions include

Table 2: Examples of oceanic data sources. DOI: https://doi.org/10.1525/elementa.418.t2

Data source	Type[a]	Content type	Ontological support	Searchable parameters (excl. key words)
ARGO	R	Float	No	Date, geo-coordinates
BCO-DMO	R	Underway, cast, float	No	Date, geo-coordinates
COPERNICUS	P	2D/3D images, cast, float	No	Date, geo-region, parameter name
EDMED	R	Underway, cast, float	Yes	Date, geo-region, geo-coordinates, parameter (ontology), instrument (ontology)
Global DMS	R	Underway	No	Date, geo-coordinates
Google dataset search	P	All	No	None
IsraMar	R	Cast	No	Date, geo-coordinates, parameter name
NCEI LAS	R	Cast, underway, 2D image, radar, float	No	Date, geo-coordinates
PANGAEA	R	Cast, underway, float	No	Date, geo-coordinates, geo-region, instrument
SeaBass	R	Cast, 2D image	No	Date, geo-coordinates, instrument
World ocean database	R	Cast, underway, 2D image, radar, float	Yes	Date, geo-coordinates, instrument, parameter name, bio-species (ontology)
Data One	P	All	Yes	Date, Geo-coordinates, instrument, parameter name, bio-species (taxonomy)

[a] R: data repositories. P: portals. Portals provide access to data from multiple repositories.

the NASA Semantic Web for Earth and Environmental Terminology (SWEET; Ashish, 2005), which hosted over 6000 concepts in 200 separate ontologies as recently as 2018, but since 2019 has been removed from public access. MarineTLO is a top-level ontology for the maritime domain (Tzitzikas et al., 2013) that contains information about marine species, ecosystems, and fishers. Significant among these efforts is OceanLink/GeoLink, a large-scale project that aims to improve discovery, access, and integration (**Figure 4**) of interdisciplinary data in the oceanographic domain (Narock et al., 2014). The ongoing project enables the discovery of integrated data from multiple repositories by creating an integrated knowledge discovery framework on top of those repositories. The project utilizes semantic web technologies, particularly ontology design patterns (ODPs; Gangemi, 2005) and a SPARQL endpoint (accessible at data.geolink.org/sparql) for semantic querying. Additional repositories supporting OBDA through a SPARQL endpoint are the European Directory of Marine Environmental Data (EDMED) (at https://edmed.seadatanet.org/sparql/), and the British oceanographic data centre, NERC SPARQL endpoint (at http://vocab.nerc.ac.uk/sparql/).

Although GeoLink's ontologies provide extensive coverage of the domain, they are far from complete. In some cases, publishing a repository's data in GeoLink is not possible due to missing concepts or a required but tedious schema-mapping process that the authors do not wish to undertake. In those cases, the remainder of the data not described by the GeoLink ontologies is published according to the provider's own schema (Krisnadhi et al., 2015). Specifically, some of the more fine-grained patterns are not fully described. For instance, in the marine biology domain, integrating data according to taxonomy can be very useful. Similarly, for measurements of plankton data such as biomass, integrating data according to plankton group size or kind can be beneficial. Such a taxonomic relation exists in the MarineTLO ontology and in WORMS but is missing in GeoLink. Another example is the lack of ontological representation of ocean basins and seas such as in SeaVoX (Claus et al., 2014). The GeoLink class *Place* can be related to a *PlaceType='ocean'* but no deeper hierarchical representation is supported. For example, if the discussed place is set to *'The Red Sea'* and some other data point is given with the place set to *'Gulf of Eilat'*, then the correct integration could not be made with GeoLink. Even if the ontological issues are resolved, realigning existing data with Geolink, or a combination of the existing ontologies, would require an extensive mapping effort that would benefit from AI-supported schema matching technologies.

Thus, scaling the search process by using OBDA would allow the collection of a large number of datasets already aligned by the domain ontology over the parameters used to perform the search step. However, using OBDA requires the domain ontology to cover all aspects of the data to be integrated, and all datasets in the repository/portal to be completely aligned with the ontology. As detailed above, current repositories and data portals mostly use taxonomies rather than ontologies, combining parameter and keyword search. Existing domain ontologies have limited coverage and cross-alignment. In the abscence of perfect OBDA systems, the *merge* phase is required to integrate

the different datasets with their mismatched schemas and data descriptions.

3.1.2. Link

The linking process entails connecting between studies and their datasets (and vice versa) and between datasets, which are derived from one or more other datasets. The prevalence of object identifiers such as DOI, coupled with the increasing tendency of authors and publishers to provide publicly available datasets together with submitted papers, has made this process easier. However, the linking process is still a largely manual process where researchers piece together the papers describing the data and vice versa. Furthermore, the linking process may require a finer resolution, as the following story published by Data One (Data Observation Network for Earth, 2020) exemplifies.

> "A third dataset looked particularly promising for use in a global study, but its PI had neglected to include units of measurement in the dataset. Unwilling to give up on a potentially great contribution, Eileen decided to do some detective work and pull up the associated publication, looking for any clues that might lead to a breakthrough. At long last, Eileen found a single table referencing the units for a particular column of data. With the units finally established, she worked backwards to make sense of the data – but at a cost of several hours' work."

Thus, even though the researcher had succeeded in linking the dataset to its corresponding publication, more refined work was needed to link specific parameters to their descriptions. This refined linkage can be delayed until the merge phase where the extended data descriptions can be used to better align the schemas of the integrated datasets with the domain ontology.

3.1.3. Identify

Even with the existence of DOI, in many cases, the same data may appear in several datasets by being used for several studies. Thus, researchers are required to meticulously read the data collection procedures of every study used to make sure that their data do not contain duplicate measurements and identify each dataset or even data point in a unique manner. The implicit danger of duplicates is that they can create an inherent bias in the results towards duplicated data. In oceanic repository integration, this process is further complicated by the fact that some records represent a collection of datasets that previously may have been published separately as well.

Thus, DOIs provide grounding of datasets to fixed, reliable repository mentions, and can be used for citation and referencing purposes. However, they do little to resolve issues such as data overlap, republication, and bundling that may manifest themselves when combining several datasets. Resolving duplicate datasets and overlapping data points using entity resolution (see Section 2.1) is an obvious use of AI-supported DI tools. As entity resolution tools rely on similarity comparisons, they would also be benefited by ocean-science-specific word embedding to allow semantic comparison.

3.2. Merge

Once a collection of datasets has been assembled, the *merge* phase can commence. To facilitate this process, one must create a mediated schema to which all other datasets are matched and subsequently mapped or use an ontology to which the datasets' schemas are mapped to facilitate OBDI. We divide this phase into three distinct steps, described in detail below. In the *match* step, correspondences are found between each attribute in every dataset's schema and the mediated schema/ontology. In the *map* phase, a function mapping from the semantics of the source dataset's schema to the mediated schema is constructed. In the *fuse* step, some datasets are interpolated over space/time to create a continuous and uniform space of measurements.

3.2.1. Match

In the match step, researchers align the different attributes/parameters in the dataset's schema with the mediated schema/ontology. To do so, the researcher must often consult the data descriptions of each parameter, which are either listed with the dataset in the source repository or described as part of the methods section of the accompanying paper. If an exact match cannot be found, the researcher must decide whether to disqualify the parameter or even the whole dataset from inclusion in the study or extend the mediated schema/ontology to accommodate the new dataset.

A wealth of literature and tools exist in the general database and knowledge-base domains to facilitate schema matching and ontology alignment. Among these are the use of acronym expansion (e.g., Sorrentino et al., 2010), a corpus of previously discovered correspondences (e.g., Madhavan et al., 2005), and instance information (e.g., Chen et al., 2018). However, to the best of our knowledge, none of these were applied to match ocean science dataset schemas, neither pair-wise nor to mediated schemas or ontologies. Zhou et al. (2018) proposed a complex real-world ontology alignment benchmark made on two separate GeoLink dataset ontologies. However, even this unique example attempts to automate ontology alignment and not automatically match dataset schemas against these ontologies. Furthermore, none of the existing automated schema matching and mapping tools is interoperable with the common ocean science meta-data formats. Schema matching can be supported further by AI-based information extraction technologies, such as described in Section 2.2.3, by extracting data descriptions from the research papers linked to the datasets. These data descriptions can be used to improve schema matching performance, thus utilizing this unique aspect of ocean science datasets.

3.2.2. Map

In some cases, the semantics of the data in one source are slightly different from that of the mediated schema/ontology. For example, a dataset may contain

two fields, one representing the latitude and another the longitude, while in the mediated schema, there exists a single *coordinates* field that combines the two. In other cases, the mediated schema may contain a field that represents a calculation performed over raw data, or the units of measurement may differ between sources. All of these examples, and other semantic differences, require a mapping phase where conversion functions are generated to facilitate data integration according to correspondences found in the matching step. Even more mundane, but crucial is the need to map from the source format to that of the central repository used to collect the data from the different datasets. For example, the data may be received in XML format and the repository stored in a relational database, requiring format conversion between the two. The use of OBDI facilitates conversion between fields of different datasets by using the encoded conversion logic within the ontology. Thus, for example, the concept of *Celsius* can be linked to the concept of *Fahrenheit* by a relation containing a specific bi-directional conversion function.

Together with the match step, there is a substantial need for golden-standard tasks and structured benchmarks for ocean science schema-matching and mapping tasks to enable the development and training of automated matching tools utilizing the existing ontologies and vocabularies. Word-embedding-based tools are highly dependent on the domain from which the text used to generate the embedding was collected. Currently absent, a word embedding for the ocean science domain would be an important enabler for AI-based DI tools (see Section 4). The same embedding could be used to enhance information extraction tools to supplement schema matching and mapping processes over datasets with information from their linked papers. As a foundational enabler, providing schema interoperability between the common ocean science data formats and those used by schema matching and mapping tools would open up a plethora of options for practitioners to use.

3.2.3. Fuse

In this step, researchers need to mitigate problems that emanate from differences in spatio-temporal resolution between the datasets. Thus, one dataset may include measurements of a 50-m depth in increments of 1 m, while another in increments of 10 cm. Decisions must be made on whether to aggregate upwards to lower resolutions, omit incompatible resolutions or interpolate the data to align the resolutions, or fill out missing data in some areas (e.g., in Kaplan and Lekien, 2007, due to faulty sensors). As previously mentioned, we leave the review and critical analysis of existing work in data fusion to future work.

In addition to spatio-temporal fusion, this step entails an additional effort of resolving duplicate and overlapping data points. While overlapping and duplicate datasets could possibly be identified at the *identify* step, identifying these cases at the datapoint level requires all fields to be aligned by the match and map steps. Here, again, we can use entity resolution to automate this task (see Example 2).

3.3. Evaluate and correct

After, or sometimes during, the data integration process, researchers must evaluate the integrated dataset to facilitate inclusion/exclusion decisions and to report quality and descriptive measures upon publication. The evaluation process often addresses one or more of the following issues.

3.3.1. Quality

Detecting data errors is often done using non-specific numerical and statistical tools; for example, by excluding all outliers, defined as values over two standard deviations from the mean. This step can be mostly aligned with the existing DI process of *data cleansing* (see Section 2.1). To identify, quantify, and possibly correct errors in data via interpolation, techniques appropriate to the data type (e.g., Gupta et al., 2014) should be used. Here, we refrain from performing a detailed review of the extent of AI used in these processes over ocean science data in the interest of brevity and focus.

A non-generic approach that could provide more accurate results can be obtained by reasoning over accumulated knowledge tied to the domain ontologies. For example, O'Brien et al. (2013) needed to remove individual samples of coccolithophore (a type of plankton) where the species was reported as *Thoracosphaera heimii*, as this species was reclassified out of the coccolithophore family after the original data were collected. This removal of misclassified samples could be done automatically by defining a logical rule over the global ontology. Furthermore, among the tools that can support a researcher in the process of evaluating the data quality of a given dataset, information extraction can provide substantial assistance. For example, information extraction tools can be used to extract and categorize quality control processes and pre-processing techniques used in a specific dataset and a collection of datasets from the scientific text describing them. Once extracted, this information can be attributed to the dataset, allowing researchers to employ data cleansing methods and filter out less trustworthy processes or, conversely, to select only those data points on which the required type of pre-processing was performed.

3.3.2. Coverage and bias

An important tool in the evaluation of result validity and relevance is the analysis of coverage and bias. Data are collected in different geographical regions, depths, and seasons, and using different instruments. When presenting results, one must either correct them for inherent biases, exclude under-represented partitions, or provide a list of caveats and analyses regarding the coverage and bias with respect to the general distribution over each dimension (geographical/temporal/other). The ability of an ocean scientist to make use of an AI-based integrated dataset strongly depends on accurate representation of possible biases and uncertainties associated with the DI process. This point is emphasized for the case of climate science studies, where uncertainties result from a wide range of sources, as a limited number of available measurements, especially for rare events (IPCC, 2014).

Existing portals/repositories provide mechanisms to filter by time/geo-location or map a collection of datasets over a world map. These mechanisms allow researchers to assess the coverage of their collection of datasets if they are from the same portal/repository. Evaluating coverage and bias over other dimensions, such as instruments used and bio-diversity, is dependent on the ability to perform OBDA, the coverage of the OBDA's ontology, and the extent of information extracted from the scientific description and aligned with the ontology.

3.4. Summary

Figure 4 presents an overview of how DI technologies (in purple/purple outline, middle layer) could support and scale the different steps and phases of the ocean science data integration process. However, to make these technologies work, some AI technologies and enablers are needed. These are listed in the bottom layer of the figure as trapezoids and are connected to the DI technologies which they support. Ontology-based technology features heavily, as it effectively combines the wealth of accumulated knowledge of the oceanic domain with AI-supported DI technologies. DI technologies and AI technologies/enablers that are missing today are drawn with a white background.

Table 3 presents a list of existing and missing enablers for DI in ocean science. Some of these enablers are presented in the figure, while others enable the processes in the figure. The gaps in the table are annotated with lower-case letters that are repeated in **Figure 4** where they are positioned on the DI technology they enable, on the AI technology they enable, or on the support a specific AI technology provides to a DI technique. Note that while the technologies and enablers reviewed in **Table 3** are listed by phase, some of them support multiple phases. For example, *entity resolution* is a DI technology that can be used to identify duplicate datasets prior to their integration in the *identify* step and to identify duplicate data points in a merged dataset as part of the *fuse* step.

4. Empirical evaluation: the impact of AI infrastructure

In the following section, we provide some empirical evidence to the necessity of creating the AI infrastructure required to support DI efforts in ocean science. As described in the previous sections, both AI-supported entity resolution tasks in the *discovery* phase and schema matching tasks in the *merge* phase could benefit from adding relevant information from unstructured sources accompanying the data. In Example 1, the fact that the Nitrate field represented the sum of nitrate and nitrite was mentioned in the column comments. The ability to retrieve this information from the comment, codify and align it with a domain ontology, relies on AI-software being able to recognize domain-specific information in unstructured text. Domain-specific datasets, benchmarks, and word embeddings are needed to bridge this gap (see **Table 3**). To exemplify the potential benefits of having this infrastructure in place, we train a state-of-the-art information extraction system on ocean science data descriptions and report on the performance gains on an information extraction task.

4.1. The task: extracting data descriptions using information extraction techniques

A standard information extraction task, named entity extraction (NER) aims to find entity mentions in unstructured text and map them into predefined classes. These entities can then be used to enrich automated data integration tasks such as schema matching and mapping. The classes a NER is seeking in the text can vary based on the requirement of a specific assignment. The most widely used classes are person, location, organization, and date (Jiang et al., 2016). For instance, a NER system trained to detect person, location, and organization when receiving the following text as input: "John Doe lives in New York City and works in the New York stock exchange," should identify the following named entities as output, where the named entity is denoted between brackets and the class between parentheses. [John Doe] (person) lives in [New York City] (location) and works in the [New York stock exchange] (organization). An ocean science DI applica-

Table 3: Missing and existing AI enablers for DI in ocean science. DOI: https://doi.org/10.1525/elementa.418.t3

Phase	Existing enablers	Remaining gaps
Discover	(1) Several ocean science ontologies. (2) OBDA to major dataset repositories. (3) Extensive use of DOI.	(a) Incomplete conceptual coverage of existing ontologies. (b) Incomplete conceptual alignment between ontologies. (c) Alignment of historical datasets with existing ontologies. (d) AI-based tools for creators to align their schemas with existing ontologies.
Merge	(4) An ocean science ontology alignment benchmark.	(e) Entity resolution oceanographic benchmarks for both dataset and data point levels. (f) Entity resolution tools utilizing ocean science word embeddings. (g) Ocean data format interoperability with existing tools. (h) Schema matching/mapping oceanographic benchmarks. (i) Matching and mapping tool utilizing semantics encoded in existing vocabularies and ontologies. (j) Word embedding for ocean science domains.
Evaluate	(5) Existing work on data cleansing/anomaly detection. Not reviewed in detail. (6) Geo-location mapping in data-portals	(k) Annotated datasets, tools, and benchmarks for extracting data quality and pre-processing descriptions from scientific text (l) Extension and refinement of oceanographic ontologies with respect to coverage, bias and quality queries.

tion would need to identify entities such as a measured variable (temperature, salinity), units (degrees, dbar), and devices (CTD, sonar, plankton counters).

4.2. Datasets

4.2.1. An oceanic science entity extraction dataset

To the best of our knowledge, no gold-standard annotated documents are freely available for the oceanic domain. Therefore, we created a small dataset to provide initial support to our claim for the need for an extensive standard to train and evaluate tools against. We retrieved 30 documents containing data descriptions from three data repositories: PANGEA (2020), BCO-DMO (Biological and Chemical Oceanography Data Management Office, 2020), and the European directory of marine environmental data (EDMED, 2020). Each token (usually a single word) was annotated in the IOB2 format using a standard NER annotation tool named TALEN (Mayhew and Roth, 2018). The IOB2 format is a tagging format designed for the NER task. The *B-* prefix before a class name is used to indicate that the token is at the beginning of a chunk, the *I-* prefix before a class indicates that the token is inside a chunk, while *O* represents a token that is not inside of any chunk. **Figure 5** shows an example of the IOB2 format used to annotate a data description document retrieved from EDMED. Our test data contain 1,256 sentences and 7,848 total tokens with an average of 262 tokens per document. We found 2,193 entities divided into 11 classes averaging 75.6 entities per document with an average length of 2.17 tokens per entity. The dataset is available online (Bar, 2020a).

4.2.2. An oceanographic text dataset

Word embeddings are created using a large text corpus. To test the hypothesis that specific word embedding could improve NER algorithms on the task of identifying oceanic entities in texts, we trained custom word vectors. Our training method is constructed based on the following steps. (a) Collect a large set of oceanographic papers. (b) Extract raw text from the collected oceanographic papers. (c) Train word embeddings based on the text corpus.

```
Environmental O
modification O
caused O
by O
aquaculture O
along O
the O
Portuguese B-GeoRegion
continental I-GeoRegion
coast I-GeoRegion
```

Figure 5: An example of the IOB2 annotation. In this figure the IOB2 annotation is used to identify a GeoRegion within a data description document retrieved from EDMED. Tokens marked with an O are not part of any entity. The token marked with B-GeoRegion begins the entity. The rest of the entity's tokens are marked with I-GeoRegion. DOI: https://doi.org/10.1525/elementa.418.f5

Due to overlapping terms from the oceanic domain in other closely related scientific domains such as earth science or biomedical science, we collected papers that were published in known oceanographic journals. We used the Crossref API (Lammey, 2015) to search for the DOIs of papers that appeared in oceanographic journals, such as *Ocean Science*, *Frontiers in Marine Science*, and *Aquatic Biology*.

After acquiring the relevant DOIs, we implemented a web crawler that searched for the full-text PDF version of the papers in several public repositories. The crawler mined 30,000 oceanic papers. We used the *Science Parse* (Clark and Divvala, 2015) open-source Java library to extract data from the papers. We extracted the title, abstract, and content section parts of the documents (references were excluded) into a JSON format. The raw text from the JSON file contained over 175 million tokens. This dataset is available online as well (Bar, 2020b).

4.3. Methods

The NER algorithm is a supervised ML model that is trained on annotated documents to recognize patterns identifying a token or set of tokens as a named entity and to which class it most likely belongs. For example, after seeing a large number of documents where the tokens next to the word *lives* describe a person (e.g., John Doe lives in), the ML model learns to classify these tokens as people. Using word embeddings to represent the documents on which the algorithm trains allows it to generalize its learned model so that similar words such as *resides* and *works* would be recognized as well. Furthermore, the token *John* itself is embedded into the vector space such that other people's names will be situated close to it. As described in Section 2, generating word embeddings is an unsupervised ML technique based on the co-occurrence of words in a very large text corpus.

In this evaluation, we use the Flair NER algorithm (Akbik et al., 2018), which is based on a word embedding technique as well. Unlike other models, the model employs character level tokenization rather than word-level tokenization. A sentence is converted to a sequence of characters, and through a language model, the algorithm learns the word representation. Flair uses a stacked embedding approach. The algorithm's character language model vector is concatenated with GloVe's word embeddings (Pennington et al., 2014) to form the final word vectors, thus leading to a better result. Flair produced state-of-the-art F1-scores on the CoNLL-03 general-purpose dataset collected from newspaper articles (Sang and De Meulder, 2003).

To adapt Flair and its NER algorithm to the oceanic domain, one can both retrain it (using supervised ML) on the classes of this domain and fine-tune the underlying word embeddings (using unsupervised ML) to reflect semantic relations in this domain better. In the following, we demonstrate both improvements.

4.3.1. Improving the Flair NER by retraining on an ocean science tagged dataset

Training was performed on a Gigabyte Technology server with an Intel i7-7700 8 core CPU, 64GB RAM, and Gigabyte GTX 1070 GPU running the Ubuntu 16.04.6 operating system. The empirical evaluation was performed using Flair version 0.4.1 (Zalando Research, 2019) running on python version 3.6.8, deployed as part of the Anaconda data science platform (Anaconda, 2020). We split our annotated dataset randomly into a training set comprised of 80% of the documents and a test set comprised of the remaining 20%. We then proceeded to train the Flair algorithm on the training set and test both the original Flair NER model and our retrained one on the test set.

4.3.2. Creating ocean science word embeddings

We utilized the oceanographic text corpus for training two new word embeddings. Word2Vec (Mikolov et al., 2013) with word-level embeddings and Flair's character-based forward and backwards embeddings (from now on, CBFB). The word-level embeddings were implemented using the Gensim Python library (Řehůřek and Sojka, 2010) and the CBFB embeddings with Flair. One of the known connections in oceanographic research is between a measured variable and its measured units. Although often a variable can be measured using different units, some notations are very common in the scientific literature. Similar to the King-Queen relationship stated by Mikolov et al. (2013) on general-purpose text, the oceanographic trained models were able to conclude the relationships in **Figure 6**. Recall that in general text, the vector representation of the word *king* was found to relate to the vector representing *queen* in the same manner as the vector *man* relates to the vector representing *woman*. After reviewing the ocean science research papers, the unsupervised algorithm, with no input from a domain expert, created an embedding model where, e.g., *m/s* relates to *speed* in the same manner that *PSU* (practical salinity units) relates to *salinity*. Note, that the fact that PSU has since been retired is unknown to the embedding algorithm, as it was trained on papers using this unit. Rather, this domain knowledge should be coded into an ontology to ensure that data from papers using PSU, can be handled appropriately when integrated with more modern datasets.

$$\overrightarrow{temperature} - \overrightarrow{salinity} + \overrightarrow{PSU} \approx \overrightarrow{degrees}$$
$$\overrightarrow{speed} - \overrightarrow{salinity} + \overrightarrow{PSU} \approx \overrightarrow{m/s}$$
$$\overrightarrow{pressure} - \overrightarrow{salinity} + \overrightarrow{PSU} \approx \overrightarrow{MPa}$$

Figure 6: Variable-unit analogies. The figure shows semantic relations between oceanic variables and their associated units, as found by the word embedding algorithm, with no intervention of a domain expert. Note that although salinity is a unitless variable, it was associated with PSU (practical salinity units) by the unsupervised algorithm. The relations can be read as follows: "*temperature* relates to *degrees* as *salinity* relates to *PSU*". DOI: https://doi.org/10.1525/elementa.418.f6

We trained the Flair algorithm with the same 80%–20% train-test split to detect data descriptions from unstructured data, where the word embeddings served as features for the NER algorithm. We ran the following stacked embeddings models: (a) GloVe and Flair embeddings trained on a general-purpose text that served as a baseline; (b) Word2Vec oceanographic model; (c) Flair's CBFB embeddings trained on an oceanographic corpus; (d) stacked embeddings model that was compiled of (b) (c) embeddings; and finally (e) stacked embeddings model of (a) (b) (c).

4.4. Evaluation measures

Several evaluation metrics have been offered to assess the efficacy of a NER system, where the most commonly used are based on the exact-match evaluation. A named entity that has been proposed by a NER system is considered correct only if there is an exact match of both entity boundaries and class (i.e., all tokens that should belong to the entity are correctly marked and assigned). However, the ML model we use in the first evaluation was not designed to detect ocean science classes (e.g., measure variable). As a result, we seek an exact boundary match with no consideration of the entity type. For example, if the NER system can detect the 'Mediterranean Sea' as a named entity, it will be considered a match regardless of the class (*location* in this example). If for the same sentence, the system will only detect 'sea' as a named entity, it will be considered a false match. In the second evaluation, we train all models to detect the specific class as well as extract the named entity and therefore seek an exact match of both boundary and entity type.

The measures precision, recall, and F1-score are arguably the most commonly used to aggregate and quantify the number of exact matches detected by a NER system. *Precision* is the fraction of true instances of the total number of instances predicted by the NER system as positive, while *recall* is the fraction of true instances predicted by the NER system of the total true instances in the dataset. *F1-score* is the harmonic mean of precision and recall. Their formal definitions are as follows.

Definition 1 (NER evaluation measures) *Let predicted positive (PP) be the set of named entities predicted as such by a NER algorithm. Let actual positive (AP) be the set of named entities that actually exist in the task. Let true positive (TP) be the intersection between these sets, i.e., those named entities that both exist in the task and were predicted by the NER algorithm, then Precision, Recall, and F1-score are defined as follows.*

$$Precision = \frac{TP}{PP} \tag{1}$$

$$Recall = \frac{TP}{AP} \tag{2}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

4.5. Results

The result of the first evaluation can be seen in **Table 4**. The F1 score of the original flair model on oceanic data is only 0.068. Training the same flair model on an oceanic dataset results in an F1 score of 0.738. The results of the second evaluation can be seen in **Table 5**. The best model was the stacked embeddings model that reached an F1 score of 0.679 on unstructured metadata. We remind the reader that in the first task, we require only a boundary match, while in the second, we require both boundary and class to be correct, making it substantially more difficult.

4.6. Discussion

The unmodified Flair model used in this evaluation scored a 0.932 F1 score on newswire text (Akbik et al., 2018). The same algorithm fails miserably on our task. The results of the retrained model can be considered as an immense improvement but still far from state-of-the-art results achieved on NER tasks in other domains. This result is expected due to the small number of training examples available to the supervised training algorithm. The result also highlights the need for an extensive, well defined, annotated dataset to train ML models over oceanic sciences tasks. Furthermore, the classes used to extract information should be carefully aligned with ocean science domain ontologies if they are to be used in conjunction with schema matching tools.

The oceanic embeddings allow Flair to boost its results on the harder boundary+class task from an F-1 of 0.415 to 0.679 for the best model. Here, too, a much more substantial increase is expected should we increase the amount of training data. Alternatively, we could use transfer learning from models trained on related datasets, such as scientific papers in general. Although 175 million tokens may sound impressive, the standard GloVe vectors used in general-purpose tasks are trained over 6 to 840 billion tokens (see Pennington et al., 2020, for examples).

5. Conclusions and future work

The study of the oceans relies on the extensive collection of physical, chemical, and biological data from various locations around the globe. Over the last century, numerous measurements have been performed continuously, resulting in the creation of an increasingly large amount of oceanic data. One of the significant challenges facing the ocean science community is to integrate this vast amount of data in a way that will facilitate its translation into improved understanding of oceanic processes. Addressing this challenge relies strongly on the implementation of AI technologies, which now, in the era of Big Data, are ubiquitously applied across scientific domains and disciplines.

In this paper, we have deconstructed the process of oceanic science DI and pointed to the key missing tools and underutilized information sources currently limiting its automation. We have focused on semantic AI technologies aiding the matching and mapping phases of the DI process, limiting our discussion of data fusion and data cleansing techniques, which we intend to address in future work.

The potential of implementing AI technologies to advance oceanic research calls for close collaboration between ocean and data scientists. Importantly, such collaboration should promote the formation of dedicated infrastructures to support AI efforts in ocean science, focusing on several activities that address major limitations in the current state of ocean data integration (**Table 3**):

- Develop AI-based tools for assisting ocean scientists in aligning their schema with existing ontologies when organizing their measurements in datasets.

Table 4: Performance of data description extraction using embeddings trained on general versus ocean science text. DOI: https://doi.org/10.1525/elementa.418.t4

Measurement[a]	Flair NER using news-trained embeddings	Flair NER using ocean-science-trained embeddings
Precision	0.221	0.746
Recall	0.040	0.731
F1 score	0.068	0.738

[a] In this task, a true positive result entails identifying a named entity regardless of its class.

Table 5: Comparative performance of Flair NER using oceanic word embeddings as features. DOI: https://doi.org/10.1525/elementa.418.t5

Embeddings method	P[a]	R	F1
Flair + GloVe (General-purpose)	0.547	0.335	0.415
Oceanic Word2Vec	0.659	0.541	0.594
Oceanic CBFB	0.705	0.648	0.676
Oceanic Word2Vec + Oceanic CBFB	0.705	0.604	0.650
Flair + GloVe (General-purpose) + Oceanic Word2Vec + Oceanic CBFB	0.713	0.649	0.679

[a] In this task, a true positive result is one where the algorithm correctly identifies the named entity and assigns the correct class.

- Extend and refine conceptual coverage of – and conceptual alignment between – existing ontologies, such that they are more compatible with the diverse and multidisciplinary nature of ocean science.
- Create ocean-science-specific schema matching and mapping benchmarks to accelerate the development of matching and mapping tools utilizing semantics encoded in existing vocabularies and ontologies.
- Similarly support the development of ocean-science-specific entity resolution tools by creating annotated datasets and benchmarks on both the dataset and data point level.
- Annotate datasets, and develop tools and benchmarks for the extraction and categorization of data quality and preprocessing descriptions from scientific text.
- Create large-scale word embeddings trained upon ocean science literature to accelerate the development of AI-based information extraction, entity resolution, and matching tools.

Formation of improved AI integration infrastructure based on these suggested activities will contribute importantly to our ability to share, explore, and interpret the vast amount of available oceanic data, thus substantially advancing ocean research.

Competing interests

The authors have no competing interests to declare.

Author contributions

T.S. Led sections 2 and 3, Y.L. led sections 1 and 5, and K.B. led section 4. The deconstruction of a DI process in ocean science as portrayed in Figure 4 was performed jointly by T.S. and Y.L.

References

Abedjan, Z, Chu, X, Deng, D, Fernandez, RC, Ilyas, IF, Ouzzani, M, Papotti, P, Stonebraker, M and **Tang, N.** 2016. Detecting Data Errors: Where are we and what needs to be done? *PVLDB* **9**(12): 993–1004. DOI: https://doi.org/10.14778/2994509.2994518

Akbik, A, Blythe, D and **Vollgraf, R.** 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20–26, 2018*, 1638–1649. https://www.aclweb.org/anthology/C18-1139/.

Alexe, B, ten Cate, B, Kolaitis, PG and **Tan, WC.** 2011. EIRENE: Interactive Design and Refinement of Schema Mappings via Data Examples. *PVLDB* **4**(12): 1414–1417. http://www.vldb.org/pvldb/vol4/p1414-alexe.pdf.

Anaconda. 2020. Anaconda Distribution. Retrieved Jan. 22nd, 2020. https://www.anaconda.com/distribution/.

Ashish, N. 2005. Semantic-Web Technology: Applications at NASA. In: Kalfoglou, Y, Schorlemmer, M, Sheth, A, Staab, S and Uschold, M (eds.), *Semantic Interoperability and Integration.* Dagstuhl, Germany: Internationales Begegnungsund Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany. (Dagstuhl Seminar Proceedings 04391). ISSN 1862-4405. http://drops.dagstuhl.de/opus/volltexte/2005/32.

Assmy, P, Henjes, J, Klaas, C and **Smetacek, V.** 2007. Mechanisms determining species dominance in a phytoplankton bloom induced by the iron fertilization experiment EisenEx in the Southern Ocean. *Deep-Sea Res Part I-Oceanogr Res Pap* **54**(3): 340–362. DOI: https://doi.org/10.1016/j.dsr.2006.12.005

Auer, S, Bizer, C, Kobilarov, G, Lehmann, J, Cyganiak, R and **Ives, ZG.** 2007. DBpedia: A Nucleus for a Web of Open Data. In: Aberer, K, Choi, K, Noy, NF, Allemang, D, Lee, K, Nixon, LJB, Golbeck, J, Mika, P, Maynard, D, Mizoguchi, R, Schreiber, G and Cudré-Mauroux, P, (eds.), *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11–15, 2007* **4825**: 722–735. Springer. DOI: https://doi.org/10.1007/978-3-540-76298-0_52

Bar, K. 2020a. Oceanic NER Project. Retrieved Jan. 22nd, 2020. DOI: https://doi.org/10.17605/OSF.IO/MY2NK

Bar, K. 2020b. Oceanic Data Description Extraction Project. Retrieved Jan. 22nd, 2020. DOI: https://doi.org/10.17605/OSF.IO/8VAFS

Bellahsene, Z, Bonifati, A and **Rahm, E.** (eds.). 2011. *Schema Matching and Mapping.* Berlin, Heidelberg: Springer. (Data-Centric Systems and Applications). ISBN 978-3-642-16517-7. DOI: https://doi.org/10.1007/978-3-642-16518-4

Berg, JL. 1976. Data base directions: the next steps. *ACM SIGMOD Record* **8**(2): 3–4. DOI: https://doi.org/10.1145/1041675.1041678

Berners-Lee, T and **Hendler, J.** 2001. Publishing on the semantic web. *Nature* **410**(6832): 1023–1024. DOI: https://doi.org/10.1038/35074206

Biological and Chemical Oceanography Data Management Office. 2020. Introduction to BCO-DMO. Retrieved Jan. 3rd, 2020. https://www.bcodmo.org/.

Bolukbasi, T, Chang, KW, Zou, JY, Saligrama, V and **Kalai, AT.** 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In: Lee, DD, Sugiyama, M, von Luxburg, U, Guyon, I and Garnett, R (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5–10, 2016,* pp. 4349–4357. Barcelona, Spain. http://papers.nips.cc/paper/6228-man-is-tocomputer-programmer-as-woman-is-to-homemakerdebiasing-word-embeddings.

British Oceanographic Data Centre. 2020. European Directory of Marine Environmental Data. Retrieved Jan. 3rd, 2020. https://edmed.seadatanet.org/.

Chen, Z, Jia, H, Heflin, J and **Davison, BD.** 2018. Generating Schema Labels Through Dataset Content Analysis. In *Companion Proc. of the The Web Conf. 2018*

(WWW '18), 1515–1522. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. DOI: https://doi.org/10.1145/3184558.3191601

Claramunt, C, Ray, C, Salmon, L, Camossi, E, Hadzagic, M, Jousselme, A-L, Andrienko, G, Andrienko, N, Theodoridis, Y and **Vouros, G.** 2017. Maritime data integration and analysis: recent progress and research challenges. In: Markl, V, Orlando, S, Mitschang, B, Andritsos, P, Sattler, K-U and Breß, S (eds.), *Proceedings of the 20th International Conference on Extending Database Technology, EDBT 2017, Venice, Italy, March 21–24, 2017*, 192–197. OpenProceedings.org. DOI: https://doi.org/10.5441/002/edbt.2017.18

Clark, CA and **Divvala, S.** 2015. Looking Beyond Text: Extracting Figures, Tables, and Captions from Computer Science Papers. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, January 25–26, 2015* **53**: 599–605. https://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/viewPaper/10092.

Claus, S, De Hauwere, N, Vanhoorne, B, Deckers, P, Souza Dias, F, Hernandez, F and **Mees, J.** 2014. Marine regions: towards a global standard for georeferenced marine names and boundaries. *Mar Geod* **37**(2): 99–125. DOI: https://doi.org/10.1080/01490419.2014.902881

Data Observation Network for Earth. 2020. The Patience of the Data Hunter. Retrieved Jan. 3rd, 2020. https://www.dataone.org/data-stories/patience-data-hunter.

De Uña, D, Rümmele, N, Gange, G, Schachte, P and **Stuckey, PJ.** 2018. Machine Learning and Constraint Programming for Relational-to-Ontology Schema Mapping. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*, 1277–1283. AAAI Press. DOI: https://doi.org/10.24963/ijcai.2018/178

Devlin, J, Chang, M-W, Lee, K and **Toutanova, K.** 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein, J, Doran, C and Solorio, T (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics. https://www.aclweb.org/anthology/N19-1423/.

Do, HH and **Rahm, E.** 2002. COMA – A System for Flexible Combination of Schema Matching Approaches. In *Proceedings of 28th International Conference on Very Large Data Bases, VLDB 2002, Hong Kong, August 20–23, 2002*, 610–621. Morgan Kaufmann. DOI: https://doi.org/10.1016/B978-155860869-6/50060-3

Doan, A, Domingos, PM and **Levy, AY.** 2000. Learning Source Description for Data Integration. In: Suciu, D and Vossen, G (eds.), *Proceedings of the Third International Workshop on the Web and Databases, WebDB 2000, Adam's Mark Hotel, Dallas, Texas, USA, May 18–19, 2000, in conjunction with ACM PODS/SIGMOD 2000. Informal proceedings,* 81–86. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.134.3677.

Doan, A, Madhavan, J, Domingos, PM and **Halevy, AY.** 2002. Learning to map between ontologies on the semantic web. In: Lassner, D, Roure, DD, Iyengar, A, (eds.), *Proceedings of the Eleventh International World Wide Web Conference, WWW 2002, May 7–11, 2002, Honolulu, Hawaii, USA,* 662–673. ACM. DOI: https://doi.org/10.1145/511446.511532

Dong, XL and **Rekatsinas, T.** 2018. Data Integration and Machine Learning: A Natural Synergy. In: Das, G, Jermaine, CM and Bernstein, PA (eds.), *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10–15, 2018*, 1645–1650. ACM. DOI: https://doi.org/10.1145/3183713.3197387

Dong, XL and **Srivastava, D.** 2015. *Big Data Integration.* Morgan & Claypool Publishers. DOI: https://doi.org/10.2200/S00578ED1V01Y201404DTM040

Durden, JM, Luo, JY, Alexander, H, Flanagan, AM and **Grossmann, L.** 2017. Integrating "Big Data" into aquatic ecology: challenges and opportunities. *Limnol Oceanogr Bull* **26**(4): 101–108. DOI: https://doi.org/10.1002/lob.10213

Ebraheem, M, Thirumuruganathan, S, Joty, SR, Ouzzani, M and **Tang, N.** 2018. Distributed representations of tuples for entity resolution. *PVLDB* **11**(11): 1454–1467. http://www.vldb.org/pvldb/vol11/p1454-ebraheem.pdf. DOI: https://doi.org/10.14778/3236187.3236198

Ekaputra, FJ, Sabou, M, Serral, E, Kiesling, E and **Biffl, S.** 2017. Ontology-based data integration in multi-disciplinary engineering environments: A Review. *Open Journal of Information Systems (OJIS)* **4**(1): 1–26.

Eriksen, CC, Osse, TJ, Light, RD, Wen, T, Lehman, TW, Sabin, PL, Ballard, JW and **Chiodi, AM.** 2001. Seaglider: A long-range autonomous underwater vehicle for oceanographic research. *IEEE J Ocean Eng* **26**(4): 424–436. DOI: https://doi.org/10.1109/48.972073

European Commission. 2020. Copernicus, the European Earth Observation and Monitoring Programme. Retrieved Jan. 1st, 2020. http://copernicus.eu/.

Fernandez, RC, Mansour, E, Qahtan, AA, Elmagarmid, AK, Ilyas, IF, Madden, S, Ouzzani, M, Stonebraker, M and **Tang, N.** 2018. Seeping Semantics: Linking Datasets Using Word Embeddings for Data Discovery. In *34th IEEE International Conference on Data Engineering, ICDE 2018, Paris, France, April 16–19, 2018,* 989–1000. IEEE Computer Society. DOI: https://doi.org/10.1109/ICDE.2018.00093

Field, CB, Behrenfeld, MJ, Randerson, JT and **Falkowski, P.** 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**(5374): 237–240. ISSN 0036-8075. DOI: https://doi.org/10.1126/science.281.5374.237

Freund, Y and **Mason, L.** 1999. The Alternating Decision Tree Learning Algorithm. In: *Proceedings of the Sixteenth International Conference on Machine Learning (ICML '99)*, 124–133. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 1558606122.

Froese, R and **Pauly, D.** 2020. FishBase. Retrieved Jan. 8th, 2020. https://www.fishbase.ca.

Gal, A. 2011. *Uncertain Schema Matching.* Morgan & Claypool Publishers. (Synthesis Lectures on Data Management). DOI: https://doi.org/10.2200/S00337ED1V01Y201102DTM013

Gal, A and **Sagi, T.** 2010. Tuning the ensemble selection process of schema matchers. *Inf Syst* **35**(8): 845–859. DOI: https://doi.org/10.1016/j.is.2010.04.003

Gangemi, A. 2005. Ontology Design Patterns for Semantic Web Content. In: Gil, Y, Motta, E, Benjamins, VR and Musen, MA (eds.), *The Semantic Web – ISWC 2005*, 262–276. Berlin, Heidelberg: Springer. DOI: https://doi.org/10.1007/11574620

Goodhue, DL, Wybo, MD and **Kirsch, LJ.** 1992. The impact of data integration on the costs and benefits of information systems. *MIS Q* **16**(3): 293–311. http://misq.org/the-impact-of-data-integration-onthe-costs-and-benefits-of-information-systems.html. DOI: https://doi.org/10.2307/249530

Gregory, K, Groth, P, Cousijn, H, Scharnhorst, A and **Wyatt, S.** 2019. Searching data: a review of observational data retrieval practices in selected disciplines. *J Assoc Inf Sci Tech* **70**(5): 419–432. DOI: https://doi.org/10.1002/asi.24165

Gruber, TR. 1995. Toward principles for the design of ontologies used for knowledge sharing? *Int J Hum-Comput Stud* **43**(5–6): 907–928. DOI: https://doi.org/10.1006/ijhc.1995.1081

Gubanov, MN, Stonebraker, M and **Bruckner, D.** 2014. Text and structured data fusion in data tamer at scale. In: Cruz, IF, Ferrari, E, Tao, Y, Bertino, E and Trajcevski, G (eds.), *IEEE 30th International Conference on Data Engineering, Chicago, ICDE 2014, IL, USA, March 31 – April 4, 2014,* 1258–1261. IEEE Computer Society. DOI: https://doi.org/10.1109/ICDE.2014.6816755

Guiry, MD and **Guiry, GM.** 2020. AlgaeBase. *World-wide electronic publication.* Galway: National University of Ireland. Searched on Jan. 8th, 2020. https://www.algaebase.org.

Gupta, M, Gao, J, Aggarwal, CC and **Han, J.** 2014. Outlier detection for temporal data: A survey. *IEEE Trans Knowl Data Eng* **26**(9): 2250–2267. ISSN 2326-3865. DOI: https://doi.org/10.1109/TKDE.2013.184

Halevy, AY, Norvig, P and **Pereira, F.** 2009. The unreasonable effectiveness of data. *IEEE Intell Syst* **24**(2): 8–12. DOI: https://doi.org/10.1109/MIS.2009.36

Halevy, AY, Rajaraman, A and **Ordille, JJ.** 2006. Data Integration: The Teenage Years. In: Dayal, U, Whang, K, Lomet, DB, Alonso, G, Lohman, GM, Kersten, ML, Cha, SK and Kim, Y (eds.), *Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Korea, September 12–15, 2006*, 9–16. ACM. http://dl.acm.org/citation.cfm?id=1164130.

Hammer, M and **McLeod, D.** 1979. On Database Management System Architecture. *Defense Technical Information Center*. http://www.dtic.mil/docs/citations/ADA076417.

Hartigan, JA and **Wong, MA.** 1979. Algorithm AS 136: A k-means clustering algorithm. *J R Stat Soc Ser C-Appl Stat* **28**(1): 100–108. DOI: https://doi.org/10.2307/2346830

He, B and **Chang, KC-C.** 2006. Automatic complex schema matching across Web query interfaces: A correlation mining approach. *ACM Trans Database Syst* **31**(1): 346–395. DOI: https://doi.org/10.1145/1132863.1132872

Hinton, G, Deng, L, Yu, D, Dahl, GE, Mohamed, A, Jaitly, N, Senior, A, Vanhoucke, V, Nguyen, P, Sainath, TN and **Kingsbury, B**. 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process Mag* **29**(6): 82–97. DOI: https://doi.org/10.1109/MSP.2012.2205597

Hogan, A, Zimmermann, A, Umbrich, J, Polleres, A and **Decker, S.** 2012. Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. *J Web Semant* **10**: 76–110. DOI: https://doi.org/10.1016/j.websem.2011.11.002

IPCC. 2014. *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* [Core Writing Team. In: Pachauri, RK and Meyer, LA (eds.)]. Geneva, Switzerland: IPCC. 151.

Jiang, R, Banchs, RE and **Li, H.** 2016. Evaluating and Combining Name Entity Recognition Systems. In *Proceedings of the Sixth Named Entity Workshop*, 21–27. Berlin, Germany: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/W16-2703

Jordan, MI and **Mitchell, TM.** 2015. Machine learning: Trends, perspectives, and prospects. *Science* **349**(6245): 255–60. DOI: https://doi.org/10.1126/science.aaa8415

Kaplan, A and **Haenlein, M.** 2019. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Bus Horiz* **62**(1): 15–25. DOI: https://doi.org/10.1016/j.bushor.2018.08.004

Kaplan, DM and **Lekien, F.** 2007. Spatial interpolation and filtering of surface current data based on open-boundary modal analysis. *J Geophys Res* **112**(C12): C12007. DOI: https://doi.org/10.1029/2006JC003984

Kenig, B and **Gal, A.** 2013. MFIBlocks: An effective blocking algorithm for entity resolution. *Inf Syst* **38**(6): 908–926. DOI: https://doi.org/10.1016/j.is.2012.11.008

Krisnadhi, AA, Hu, Y, Janowicz, K, Hitzler, P, Arko, RA, Carbotte, S, Chandler, C, Cheatham, M, Fils, D, Finin, T, Ji, P, Jones, MB, Karima, N, Lehnert, KA, Mickle, A, Narock, T, O'Brien, M, Raymond, L, Shepherd, A, Schildhauer, M and **Wiebe, P.**

2015. The GeoLink Framework for Pattern-based Linked Data Integration. In: Villata, S, Pan, JZ and Dragoni, M (eds.), *Proceedings of the ISWC 2015 Posters & Demonstrations Track co-located with the 14th International Semantic Web Conference (ISWC-2015), Bethlehem, PA, USA, October 11, 2015, CEUR Workshop Proceedings,* vol. 1486. CEUR-WS.org. (CEUR Workshop Proceedings, vol. 1486). http://ceur-ws.org/Vol-1486/paper_99.pdf.

Krizhevsky, A, Sutskever, I and **Hinton, GE.** 2017. ImageNet classification with deep convolutional neural networks. *Commun ACM* **60**(6): 84–90. DOI: https://doi.org/10.1145/3065386

Lammey, R. 2015. CrossRef Text and Data Mining Services. *Science Editing* **2**: 22–27. DOI: https://doi.org/10.6087/kcse.32

Leadbetter, A, Hamre, T, Lowry, R, Lassoued, Y and **Dunne, D.** 2010. Ontologies and ontology extension for marine environmental information systems. In: Arne, J, Berre, DR, Maue, P, (eds.), *Proceedings of the Workshop Environmental Information Systems and Services-Infastructures and Platforms, (envip'2010), Bonn, Germany* **34**(25): 12–14. http://ceur-ws.org/Vol-679/paper11.pdf.

Leblanc, K, Arístegui, J, Armand, L, Assmy, P, Beker, B, Bode, A, Breton, E, Cornet, V, Gibson, J, Gosselin, MP, Kopczynska, E, Marshall, H, Peloquin, J, Piontkovski, S, Poulton, AJ, Quéguiner, B, Schiebel, R, Shipe, R, Stefels, J, van Leeuwe, MA, Varela, M, Widdicombe, C and **Yallop, M.** 2012. A global diatom database – abundance, biovolume and biomass in the world ocean. *Earth Syst Sci Data* **4**: 149–165. DOI: https://doi.org/10.5194/essd-4-149-2012

Lehahn, Y, d'Ovidio, F and **Koren, I.** 2018. A satellite-based lagrangian view on phytoplankton dynamics. *Annu Rev Mar Sci* **10**: 99–119. DOI: https://doi.org/10.1146/annurev-marine-121916-063204

Lehahn, Y, Ingle, KN and **Golberg, A.** 2016. Global potential of offshore and shallow waters macroalgal biorefineries to provide for food, chemicals and energy: feasibility and sustainability. *Algal Res* **17**: 150–160. DOI: https://doi.org/10.1016/j.algal.2016.03.031

Lesiv, M, Moltchanova, E, Schepaschenko, D, See, L, Shvidenko, A, Comber, A and **Fritz, S.** 2016. Comparison of data fusion methods using crowdsourced data in creating a hybrid forest cover map. *Remote Sens* **8**(3): 261. DOI: https://doi.org/10.3390/rs8030261

Levy, O, Goldberg, Y and **Dagan, I.** 2015. Improving distributional similarity with lessons learned from word embeddings. *TACL* **3**: 211–225. DOI: https://doi.org/10.1162/tacl_a_00134

Lumpkin, R, Özgökmen, T and **Centurioni, L.** 2017. Advances in the application of surface drifters. *Annu Rev Mar Sci* **9**: 59–81. DOI: https://doi.org/10.1146/annurev-marine-010816-060641

Luo, Y-W, Doney, SC, Anderson, LA, Benavides, M, Berman-Frank, I, Bode, A, Bonnet, S, Boström, KH, Böttjer, D, Capone, DG, Carpenter, EJ, Chen, YL, Church, MJ, Dore, JE, Falcón, LI, Fernández, A, Foster, RA, Furuya, K, Gómez, F, Gundersen, K, Hynes, AM, Karl, DM, Kitajima, S, Langlois, RJ, LaRoche, J, Letelier, RM, Marañón, E, McGillicuddy, DJ, Moisander, PH, Moore, CM, Mouriño-Carballido, B, Mulholland, MR, Needoba, JA, Orcutt, KM, Poulton, AJ, Rahav, E, Raimbault, P, Rees, AP, Riemann, L, Shiozaki, T, Subramaniam, A, Tyrrell, T, Turk-Kubo, KA, Varela, M, Villareal, TA, Webb, EA, White, AE, Wu, J and **Zehr, JP.** 2012. Database of diazotrophs in global ocean: abundances, biomass and nitrogen fixation rates. *Earth Syst Sci Data* **4**: 47–73. DOI: https://doi.org/10.5194/essd-4-47-2012

Madhavan, J, Bernstein, PA, Doan, A and **Halevy, AY.** 2005. Corpus-based Schema Matching. In: Aberer, K, Franklin, MJ and Nishio, S (eds.), *Proceedings of the 21st International Conference on Data Engineering, ICDE 2005, 5–8 April 2005, Tokyo, Japan,* 57–68. IEEE Computer Society. DOI: https://doi.org/10.1109/ICDE.2005.39

Martinez-Rodriguez, JL, Hogan, A and **Lopez-Arevalo, I.** 2020. Information extraction meets the semantic web: a survey. *Semant Web* **11**(2): 255–335. DOI: https://doi.org/10.3233/SW-180333

Mayhew, S and **Roth, D.** 2018. TALEN: Tool for Annotation of Low-resource ENtities. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, Melbourne, Australia, July 15–20, 2018*, 80–86. DOI: https://doi.org/10.18653/v1/P18-4014

Mikolov, T, Yih, W and **Zweig, G.** 2013. Linguistic Regularities in Continuous Space Word Representations. In: Vanderwende, L, III HD and Kirchhoff, K (eds.), *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9–14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA,* 746–751. The Association for Computational Linguistics. https://www.aclweb.org/anthology/N13-1090/.

Narock, T, Arko, RA, Carbotte, S, Krisnadhi, A, Hitzler, P, Cheatham, M, Shepherd, A, Chandler, C, Raymond, L, Wiebe, P and **Finin, TW.** 2014. The OceanLink project. In: Lin, JJ, Pei, J, Hu, X, Chang, W, Nambiar, R, Aggarwal, CC, Cercone, N, Honavar, VG, Huan, J, Mobasher, B and Pyne, S (eds.), *2014 IEEE International Conference on Big Data, Big Data 2014, Washington, DC, USA, October 27–30, 2014*, 14–21. IEEE Computer Society. DOI: https://doi.org/10.1109/BigData.2014.7004347

National Oceanic and Atmospheric Administration. 2020a. Big Data Project. Retrieved Jan. 3rd, 2020. https://www.noaa.gov/big-dataproject.

National Oceanic and Atmospheric Administration. 2020b. National Centers for Environmental Information. Retrieved Jan. 1st, 2020. https://www.ncei.noaa.gov/.

Nickel, M, Murphy, K, Tresp, V and **Gabrilovich, E.** 2016. A review of relational machine learning for

knowledge graphs. *Proc IEEE* **104**(1): 11–33. DOI: https://doi.org/10.1109/JPROC.2015.2483592

O'Brien, CJ, Peloquin, JA, Vogt, M, Heinle, M, Gruber, N, Ajani, P, Andruleit, H, Arístegui, J, Beaufort, L, Estrada, M, Karentz, D, Kopczyńska, E, Lee, R, Poulton, AJ, Pritchard, T and **Widdicombe, C.** 2013. Global marine plankton functional type biomass distributions: Coccolithophores. *Earth Syst Sci Data* **5**(2): 259–276. DOI: https://doi.org/10.5194/essd-5-259-2013

O'Hare, K, Jurek-Loughrey, A and **de Campos, C.** 2019. A Review of Unsupervised and Semi-supervised Blocking Methods for Record Linkage. In: Deepak, P and Jurek-Loughrey, A (eds.), *Linking and Mining Heterogeneous and Multi-view Data*, 79–105. Cham: Springer International Publishing: ISBN 978-3-030-01872-6. DOI: https://doi.org/10.1007/978-3-030-01872-6

PANGEA. 2020. PANGEA, Data Publisher for Earth and Environmental Science. Retrieved Jan. 1st, 2020. https://pangaea.de/.

Papadakis, G, Svirsky, J, Gal, A and **Palpanas, T.** 2016. Comparative analysis of approximate blocking techniques for entity resolution. *PVLDB* **9**(9): 684–695. DOI: https://doi.org/10.14778/2947618.2947624

Pennington, J, Socher, R and **Manning, CD.** 2014. Glove: Global Vectors for Word Representation. In: Moschitti, A, Pang, B and Daelemans, W (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 1532–1543. ACL. https://www.aclweb.org/anthology/D14-1162/. DOI: https://doi.org/10.3115/v1/D14-1162

Pennington, J, Socher, R and **Manning, CD.** 2020. GloVe: Global Vectors for Word Representation. Retrieved Jan. 22nd, 2020. https://nlp.stanford.edu/projects/glove/.

Prud'hommeaux, E and **Seaborne, A.** 2008. SPARQL Query Language for RDF.W3C. http://www.w3.org/TR/rdf-sparql-query/.

Řehůřek, R and **Sojka, P.** 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 46–50. Valletta, Malta: ELRA. http://is.muni.cz/publication/884893/en.

Roemmich, D, Johnson, GC, Riser, S, Davis, R, Gilson, J, Owens, WB, Garzoli, SL, Schmid, C and **Ignaszewski, M.** 2009. The Argo Program: Observing the global ocean with profiling floats. *Oceanogr* **22**: 34–43. DOI: https://doi.org/10.5670/oceanog.2009.36

Russakovsky, O, Deng, J, Su, H, Krause, J, Satheesh, S, Ma, S, Huang, Z, Karpathy, A, Khosla, A, Bernstein, MS, Berg, AC and **Li, F.** 2015. ImageNet large scale visual recognition challenge. *Int J Comput Vis* **115**(3): 211–252. DOI: https://doi.org/10.1007/s11263-015-0816-y

Sagi, T and **Gal, A.** 2013. Schema matching prediction with applications to data source discovery and dynamic ensembling. *VLDB J* **22**(5): 689–710. DOI: https://doi.org/10.1007/s00778-013-0325-y

Sagi, T, Gal, A, Barkol, O, Bergman, R and **Avram, A.** 2017. Multi-source uncertain entity resolution: transforming holocaust victim reports into people. *Inf Syst* **65**: 124–136. DOI: https://doi.org/10.1016/j.is.2016.12.003

Sang, EFTK and **De Meulder, F.** 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Daelemans, W and Osborne, M (eds.), *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 – June 1, 2003*, 142–147. ACL. https://www.aclweb.org/anthology/W03-0419/.

Schmidt, EM and **Kim, YE.** 2011. Learning emotion-based acoustic features with deep belief networks. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2011, New Paltz, NY, USA, October 16–19, 2011*, 65–68. IEEE. DOI: https://doi.org/10.1109/ASPAA.2011.6082328

Semina, GI and **Mikaelyan, AS.** 1994. (Table 1) Hydrological, hydrooptical, and hydrochemical characteristics of seawater at 7 stations in the Northwest Pacific. PANGAEA. In supplement to: Semina, GI; Mikaelyan, AS (1994): Phytoplankton of various size groups from the Northwest Pacific Ocean during summer. *Oceanology* **33**(5): 618–624. DOI: https://doi.org/10.1594/PANGAEA.759517

Shvaiko, P and **Euzenat, J.** 2013. Ontology matching: state of the art and future challenges. *IEEE Trans Knowl Data Eng* **25**(1): 158–176. DOI: https://doi.org/10.1109/TKDE.2011.253

Silver, D, Huang, A, Maddison, CJ, Guez, A, Sifre, L, van den Driessche, G, Schrittwieser, J, Antonoglou, I, Panneershelvam, V, Lanctot, M, Dieleman, S, Grewe, D, Nham, J, Kalchbrenner, N, Sutskever, I, Lillicrap, TP, Leach, M, Kavukcuoglu, K, Graepel, T and **Hassabis, D.** 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**(7587): 484–489. DOI: https://doi.org/10.1038/nature16961

Sorrentino, S, Bergamaschi, S, Gawinecki, M and **Po, L.** 2010. Schema label normalization for improving schema matching. *Data Knowl Eng* **69**(12): 1254–1273. DOI: https://doi.org/10.1016/j.datak.2010.10.004

Stocker, TF, Qin, D, Plattner, GK, Alexander, LV, Allen, SK, Bindoff, NL, Bréon, FM, Church, JA, Cubasch, U, Emori, S, Forster, P, Friedlingstein, P, Gillett, N, Gregory, JM, Hartmann, DL, Jansen, E, Kirtman, B, Knutti, R, Krishna Kumar, K, Lemke, P, Marotzke, J, Masson-Delmotte, V, Meehl, GA, Mokhov, II, Piao, S, Ramaswamy, V, Randall, D, Rhein, M, Rojas, M, Sabine, C, Shindell, D, Talley, LD, Vaughan, DG and **Xie, SP.** 2013. Technical Summary. In: Stocker, T, Qin, D, Plattner, GK, Tignor, M, Allen, S, Boschung, J, Nauels, A, Xia, Y, Bex, V and Midgley, P (eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Inter-governmental*

Panel on Climate Change. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Tzitzikas, Y, Allocca, C, Bekiari, C, Marketakis, Y, Fafalios, P, Doerr, M, Minadakis, N, Patkos, T and **Candela, L.** 2013. Integrating Heterogeneous and Distributed Information about Marine Species through a Top Level Ontology. In: Garoufallou, E and Greenberg, J (eds.), *Metadata and Semantics Research – 7th Research Conference, MTSR 2013, Thessaloniki, Greece, November 19–22, 2013. Proceedings, Communications in Computer and Information Science* **90**: 289–301. Springer. (Communications in Computer and Information Science, vol. 390). DOI: https://doi.org/10.1007/978-3-319-03437-9_29

UNIDATA. 2019. Network Common Data Form (NetCDF). Retrieved Jan. 3rd, 2020. https://www.unidata.ucar.edu/software/netcdf/.

Uschold, M. 1998. Knowledge level modelling: concepts and terminology. *Knowl Eng Rev* **13**(1): 5–29. DOI: https://doi.org/10.1017/S0269888998001040

Voyant, C, Notton, G, Kalogirou, S, Nivet, M, Paoli, C, Motte, F and **Fouilloy, A.** 2017. Machine learning methods for solar radiation forecasting: A review. *Renew Energy* **105**: 569–582. DOI: https://doi.org/10.1016/j.renene.2016.12.095

Waltz, E and **Waltz, T.** 2017. Principles and practice of image and spatial data fusion. In: Liggins, M, II, Hall, D and Llinas, J (eds.), *Handbook of multisensor data fusion,* 109–134. CRC Press: DOI: https://doi.org/10.1201/9781420053098

Wang, X, Xu, J, Liu, M, Wei, Z, Bu, W and **Hong, T.** 2017. An ontology-based approach for marine geochemical data interoperation. *IEEE Access* **5**: 13364–13371. DOI: https://doi.org/10.1109/ACCESS.2017.2724641

WoRMS Editorial Board. 2020. World Register of Marine Species (WoRMS). Accessed: 2020-01-03. http://www.marinespecies.org.

Xiao, G, Calvanese, D, Kontchakov, R, Lembo, D, Poggi, A, Rosati, R and **Zakharyaschev, M.** 2018. Ontology-Based Data Access: A Survey. In: Lang, J, (ed.), *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13–19, 2018, Stockholm, Sweden*, 5511–5519. ijcai.org. DOI: https://doi.org/10.24963/ijcai.2018/777

Zalando Research. 2019. flair: A very simple framework for state-of-the-art NLP. Retrieved March 21st, 2020. https://github.com/flairNLP/flair.

Zeng, ML. 2008. Knowledge Organization Systems (KOS). *Knowl Organ* **35**(2–3): 160–182. DOI: https://doi.org/10.5771/0943-7444-2008-2-3-160

Zhou, L, Cheatham, M, Krisnadhi, A and **Hitzler, P.** 2018. A Complex Alignment Benchmark: GeoLink Dataset. In: Vrandečić, D, Bontcheva, K, Suárez-Figueroa, MC, Presutti, V, Celino, I, Sabou, M, Kaffee, L-A and Simperl, E, (eds.), *The Semantic Web – ISWC 2018 – 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II* **11137**: 273–288. Springer. DOI: https://doi.org/10.1007/978-3-030-00668-6n_17

How to cite this article: Sagi, T, Lehahn, Y and Bar, K. 2020. Artificial intelligence for ocean science data integration: current state, gaps, and way forward. *Elem Sci Anth,* 8: 21. DOI: https://doi.org/10.1525/elementa.418

Domain Editor-in-Chief: Jody W. Deming, School of Oceanography, University of Washington, US

Associate Editor: Lisa A. Miller, Institute of Ocean Sciences, Fisheries and Oceans Canada, CA

Knowledge Domain: Ocean Science

Submitted: 21 August 2019 **Accepted:** 22 April 2020 **Published:** 15 May 2020

ELEMENTA Science of the Anthropocene

Elem Sci Anth is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS