Environmental and life sciences observations in knowledge graphs using NLP techniques to support multidisciplinary studies

Muhammad Arslan, Jean-Christophe Desconnets, Isabelle Mougenot

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2022) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

The 5th International Conference on Emerging Data and Industry 4.0 (EDI40)
March 22-25, 2022, Porto, Portugal

Environmental and life sciences observations in knowledge graphs using NLP techniques to support multidisciplinary studies

Muhammad Arslan[a,] *, Jean-Christophe Desconnets[a], Isabelle Mougenot[b]

[a]*Mission Infrastructures et Données Numériques, IRD, F-13572 Marseille Cedex 02, France.*
[b]*UMR 228 Espace Dev UM, Maison de la Teledetection, 500 rue JF Breton, 34093 Montpellier Cedex 5, France*

Abstract

The understanding of environmental observations is a continuous challenge for environmental and life science investigations. The environmental data is complex as it involves its own features, methods, properties, systems, and spatio-temporal dimensions. The time granularity remains approximately the same for different environmental contexts but geographic and rest of the above-mentioned entities are defined using domain vocabularies that are specific for each discipline. It is time-consuming for the researchers of life sciences` discipline to discover, access, and analyze relevant environmental observations as each discipline has its data formats, vocabularies, and metadata standards. These differences introduce structural and semantic heterogeneities, resulting in creating a barrier for reusing datasets generated by other disciplines. Existing dataset discovery platforms contain domain-specific metadata descriptions for explaining datasets which limits their usage. To overcome this knowledge barrier, this work reports the proof-of-concept implementation of a knowledge graph that is centered towards the oceanography use case scenario using NLP techniques (named entity recognition (NER) followed by text preprocessing). The constructed knowledge graph is a collection of subgraphs each representing the metadata of a dataset. It uses the geo-spatial and open semantic data standards that aim to provide enhanced metadata descriptions of datasets for enabling multidisciplinary research.

Keywords: Metadata; user needs; observations; life sciences; datasets; knowledge graph; named-entity recognition; NLP

* Corresponding author. Tel.: +33 46 74 16 100
E-mail address: muhammad.arslan@ird.fr

1. Introduction

A major scientific challenge faced by the global change researchers is first to understand the environment and its linked processes, and then to establish the correlations with the evolution of living organisms (i.e. a study of life sciences) using the acquired knowledge [1, 2]. The study of environmental science is complex having many subdomains such as Ecology, Biology, Plant Science, Zoology, and Oceanography, and each subdomain has its own vocabularies, and methodologies for data discovery and access [3]. These differences result in structural and semantic heterogeneities within a dataset as well as in metadata (i.e. data describing a dataset) representation making it difficult for researchers of other disciplines to find, combine or integrate data of different environmental observations [3].

To help overcome this problem of data representation of scientific observations, open standards, and Semantic Web technologies can transform dataset descriptions (i.e. metadata) into coherent and interoperable infrastructures [3, 4]. This process needs the construction of knowledge graphs (i.e. graphs of data) for accumulating and conveying knowledge of the real world, whose nodes represent entities and whose edges represent relationships between these entities [5]. “Structuring data through an ontology into a knowledge graph opens up a unifying horizon of meaning for the interlinked entities and also new multidisciplinary studies” [5]. From the existing studies, for developing the knowledge graph, the Sensor, Observation, Sample, and Actuator ontology (SOSA) was chosen as it provides a flexible and lightweight vocabulary for representing the characteristics of environmental observations [4]. Using the SOSA ontology as the core, along with semantic open data standards, including Resource Description Framework (RDF), RDF Schema (RDFS), and Web Ontology Language (OWL), this work discusses the implementation of a knowledge graph using the oceanography case study. We propose to demonstrate the suitability of the knowledge graph to improve the characterization of datasets by extending their technical descriptions so that dataset discovery processes are improved.

The rest of the paper is organized as follows: Section 2 describes the background of the study. Section 3 discusses the knowledge graph using the oceanography case study for improving dataset discovery processes. Section 4 presents the implementation of the knowledge graph. Section 5 contains a brief discussion, and section 6 presents a conclusion.

2. Background

The literature review is done to discover existing systems of data and associated metadata discovery for environmental and life sciences contexts. A couple of most recent data discovery models [6, 7] are proposed based on Relational Database Management Systems (RDMS). Also, the implementation of Semantic Web technologies for representation and storage of environmental and life sciences datasets [8 – 12] are investigated using graph-based data. The core of data representations for constructing these systems based on Semantic Web technologies has been the Resource Description Framework (RDF). SPARQL (Simple Protocol and RDF Query Language) Protocol – a World Wide Web Consortium (W3C) standard – is used for querying the RDF-based datastores [9, 10]. Furthermore, the domain ontologies and data vocabularies are constructed to foster data interoperability through the use of RDF language, RDFS and SPARQL-based querying among disconnected datasets [13]. Beyond the specialized ontologies, mid-level ontologies that can be used for environmental and life sciences applications are also built. For example, SOSA and Extended-SOSA ontology are constructed for storing environmental observations [4].

To overcome the issue of data representation and discovery of environmental observations, several systems as mentioned above are developed. However, these systems contain specialized metadata descriptions that can be difficult to use by non-experts interested in conducting multidisciplinary analyses. To address this research gap, a knowledge graph is implemented that is aligned with linked data, and the W3C technology stack aimed at forming a global interconnected data graph [5]. The fundamental requirements for constructing a knowledge graph were chosen using the existing literature [7, 14, 15]. These requirements were primarily based on 5 different types of information about; 1) the studied feature and its acquisition, 2) spatial dimension (i.e. data about location), 3) temporal dimension (e.g. time granularity), 4) dataset ownership, and 5) dataset access details. These requirements were translated into 9 questions for developing a knowledge graph (Table 1). The knowledge graph should be based

on Semantic Web-based approaches that offer an open data infrastructure [13] through the reuse of existing ontologies, and enable us to make domain-neutral dataset searches.

Table 1. Use of existing ontologies and vocabularies for defining the concepts involved in requirement questions.

Requirement questions	Type	Concepts	Ontologies and vocabularies used
Q1. Which real-world feature was monitored?	Feature and its acquisition	FOI	SOSA
Q2. What property of a feature was observed?		Observable property	Complex Property Model (CPM)
Q3. How was the property data monitored?		Procedure	SOSA
Q4. In what units are the collected data expressed?		Unit of measurement	CPM
Q5. How was the dataset captured?		Sensor and platform	SOSA
Q6. Where exactly the property was measured?	Spatial	Spatial representation	SWEET and GeoNames
Q7. When was the data collected?	Temporal	Temporal representation	SWEET and Time
Q8. Who collected the data?	Ownership	Contributor	Data Catalog Vocabulary (DCAT)
Q9. Where can the data be accessed?	Data access	DownloadURL	DCAT

3. Knowledge graph modeling using a case study

After an extensive literature review, a list of competency questions was drawn up, and the concepts from the existing ontologies and data vocabularies are selected (see Table 1) that can be used for the knowledge graph. The reason of such selection is straightforward since they are compliant with the Open Geospatial Consortium (OGC) providing an open semantic data infrastructure [13]. Our knowledge graph [16] is based on a SOSA with its extension [4, 17] (see Fig. 1), to define the metadata for a dataset. Each observation links to: 1) a *sensor*, 2) an *ObservableProperty*, and 3) an *FOI* for detailing the property with which it was linked. *Observations* are combined into collections (*ObservationCollection*) to share the same properties [4, 17]. To this end, an *ObservableProperty* concept is defined from the SOSA ontology using Complex Properties Model (CPM) ontology [18] to add more details about the examined observed property using subproperties (i.e. defined as *object of interest* using CPM ontology) each having statistical measure (i.e. *Unit*)). To specify the dataset occurrence properties, we use the DCAT ontology [19] and DCTERMS [20]. To add the spatial dataset granularity, the class named "*representation (REPR)*" from the SWEET ontology [21] is used for providing spatio-temporal dimensions. In addition, the graph uses GEONAMES to specify the spatial features, and OWL-Time [22] for describing temporal concepts. In addition, the foundations of the W3C Web semantic ontology were also utilized, i.e. RDFS [23] and SKOS [24] ontologies.

To demonstrate the application of our knowledge graph, metadata is added to it (see Fig. 1) by using the SWEET ontology, Geonames gazeeter and AGROVOC thesaurus to define the concepts pertaining to the selected multidisciplinary research area (in this case, centered around oceanography). The dataset (REPHY Monitoring Network data) was retrieved from the online data repository of Ifremer [25]. Data of two acquired parameters were considered from this dataset which are; 1) the detection and counting of phytoplankton species and, 2) the physico-chemical property of water bodies. As shown in Fig. 1, the REPHY monitoring network data collection (at level 0) consists of two separate *ObservationCollections* (*PhytoplanktonPresenceObs* and *Physico-chemicalParaObs*). Each member of the REPHY data collection shares the same *FOI*, i.e. *CoastalWater*, defined using the Office International de l'Eau (OIEAU) ontology. Each *ObservationCollection* corresponds to a different dataset defined using a *sosa:hasResult* property. Every *Observation* that is a member of *PhytoplanktonPresenceObs* will share the same properties that are assigned to the *PhytoplanktonPresenceObs* collection (at level 1), such as *sosa:hasResult*. For the *LicmophoraPresenceObs* observation, the *Count* concept is set as the *ObservableProperty* using the SWEET ontology. The *ObservableProperty* is further defined using CPM to add more details about the examined property (*Count*) by *ObjectOfInterest* and *Property*. Furthermore, the *ObjectOfInterest* is further enriched with *Phytoplankton*, and *Plankton* concepts defined by the AGROVOC vocabulary using SKOS relations. The *Physico-chemicalParaObs* collection (at level 1 in Fig. 1) contains observations of *Temperature* readings. In this case, the observations were tagged to a *Sensor* concept and later labelled with its tag (*Temperature sensor in-situ*). In addition, its *ObservableProperty* is defined using a *Temperature* concept by the SWEET ontology. Moreover, the dataset is enriched with identification metadata information such as *title*, *contributor*, *data distribution information*, *date of issue*, *publisher details*, and *keywords* using DCAT vocabulary [19, 20]. These supplementary annotations provide general and technical descriptions useful for evaluating whether a dataset is suitable for a particular purpose

and for accessing it.

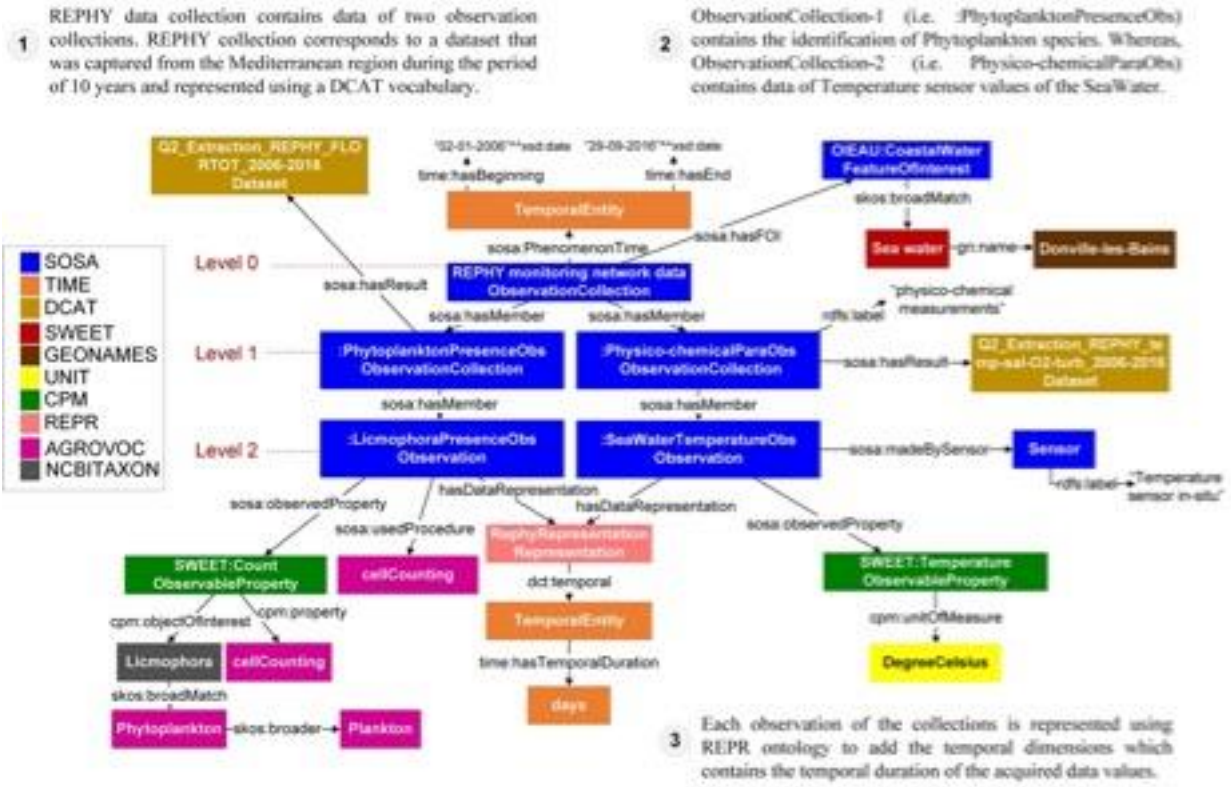

Fig. 1. Knowledge graph for the oceanography example

Prefix namespaces: **UNIT**: http://qudt.org/2.1/vocab/unit/; **SKOS**: http://www.w3.org/2004/02/skos/core#; **OIEAU**: https://www.oieau.fr/; **SWEET**: http://sweetontology.net/; **GEONAMES**: http://www.geonames.org/ontology#; **SOSA**: http://www.w3.org/ns/sosa/; **NCBITAXON**: https://www.ncbi.nlm.nih.gov/taxonomy/; **AGROVOC**: http://aims.fao.org/aos/agrovoc/

4. Proof of concept system

The knowledge graph as discussed in Section 3 is implemented using the Python programming language. For the prototype development, open semantic data standards of the W3C and named entity recognition (NER) i.e. a form of Natural Language Processing (NLP) are used. Concepts used for defining metadata are recognized using their unique Uniform Resource Identifiers (URIs). The concepts and their mappings are stored as RDF triples in the system. These RDF triples reside in the in-memory store to support SPARQL queries later. For showing the usefulness of the system, its execution, which encompasses 3 modules is shown in Fig. 2.

The 1st module involves extracting the metadata fields of a dataset. This process is executed using two Python libraries, which are; *BeautifulSoap* [26] and *Requests* [27]. *BeautifulSoap* is used for pulling data out of HTML pages whereas, *Requests* allows us to send HTTP requests to HTML pages to download data. Using these two libraries and by providing the *dataset URI*, *dataset title*, *description*, *published date*, *publisher*, *temporal duration*, *distribution name*, *distribution URL*, *spatial coverage*, *beginning date and end date of a dataset* are extracted. Out of all these extracted fields, only *dataset title* and *description* are processed using the *Spacy* Python library [28] to extract named entities such as *dates*, *locations*, *FOIs*, *the object of interests* and *properties*. Within the *Spacy*, the *PhraseMatcher* [28] is used to match the concepts that exist in *dataset title* and *description* with large terminology lists. These lists are created based on AGROVOC controlled vocabulary containing the categorization (e.g. *features*, *methods*, *properties* and *systems*) of concepts related to environmental and life sciences. Using the *PhraseMatcher*, the below-mentioned possible labelling is achieved which can help the user to select the most appropriate category to fill metadata fields of a graph (see Fig. 3).

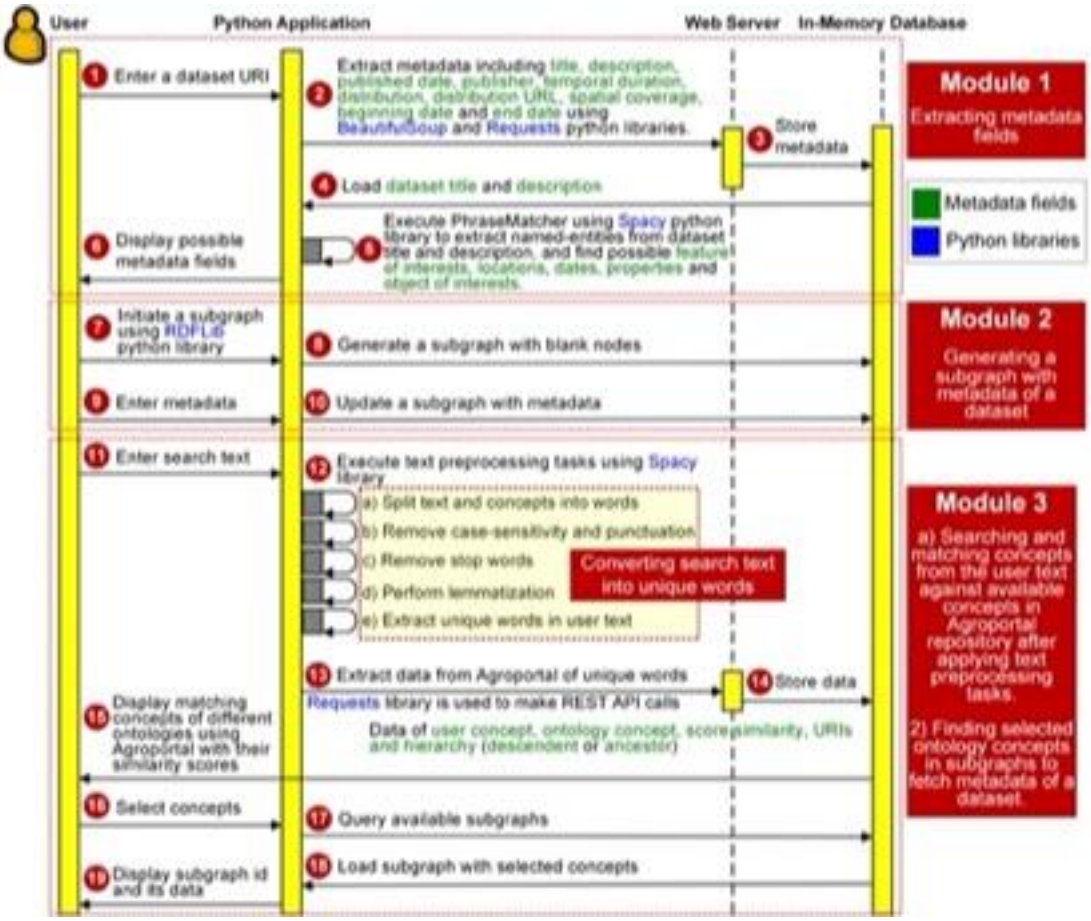

Fig. 2. Sequence diagram of implementing different modules to construct a knowledge graph.

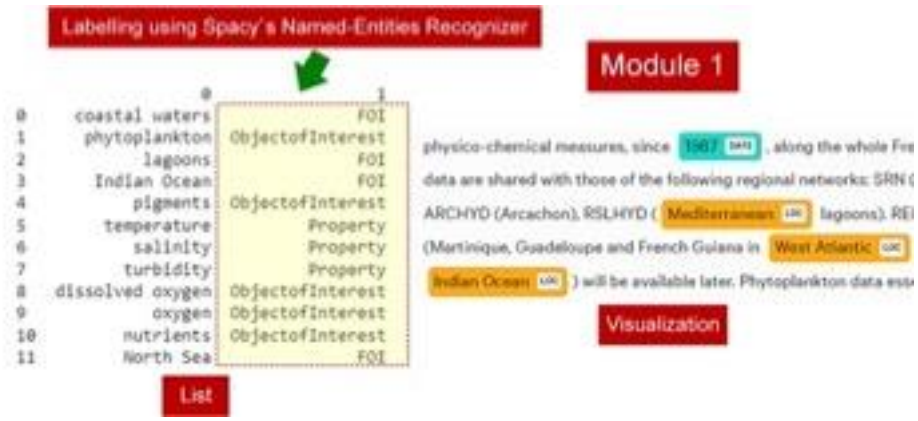

Fig. 3. Labelling different concepts from a dataset title and description using the NER

The 2nd module involves the generation of subgraphs with blank nodes and later these nodes are filled using the metadata fields extracted in module 1. This is achieved using the *RDFLib* [29] Python library providing a simple and powerful language for working with RDF and representing information. Using this library, subgraphs are created using metadata information based on the user input, as shown in Fig. 4. After creating the subgraphs, the user can query the subgraphs to extract the stored information using module 3. As the user enters the text which

needs to be searched in subgraphs, the text is broken down into discrete words using text preprocessing tasks [28]. The *case sensitivity*, *punctuation* and *stop words* (e.g. “a”, “the”, “is”, “are”, etc.) are removed. Later, the *lemmatization* technique is executed for transforming words to their normalized forms (i.e. a root format e.g. changing to change) using a dictionary. Last, the unique words are extracted from the preprocessed words. These unique words (also called concepts) are searched over the AgroPortal (http://agroportal.lirmm.fr/). The AgroPortal contains the concepts with their ancestors and descendants’ concepts are defined using reference ontologies and vocabularies related to agronomy. The AgroPortal data is extracted using the REST API offered by the *Requests* Python library. The parameter of *similarity score* is used to match the user concepts with the ontology concepts. The score is computed using the *Levenshtein edit-distance* between two words [28]. The *edit distance* is the number of characters that need to be changed for transforming word1 into word2 [28]. After calculating the *similarity scores*, the concepts with high scores are displayed to the user. Based on the user selection of concepts, the existing subgraphs are queried for finding the selected concepts or their ancestors and descendants. Upon matching the concepts, the relevant subgraph name and its related information to access the actual dataset will be displayed to the user, as shown in Fig. 4.

Fig. 4. Graphical User Interface (GUI) of constructing a knowledge graph and querying it using the user text.

4. Discussion

The implementation of a knowledge graph is based on the Semantic Web and text processing methods. Its development involves 3 modules, which are; 1) extracting metadata fields [30] using the dataset URI, 2) subgraphs generation using extracted metadata fields got from module 1, and 3) a user search module which includes; a) transforming the user text into unique concepts using text preprocessing techniques, b) searching the unique concepts over the AgroPortal for extracting their hierarchies (i.e. transitive concepts including ancestors and descendants), and c) searching the user-selected concepts in the existing subgraphs. Once the relevant subgraph is

displayed to the user, its organized metadata using observations can be studied. Each observation links to 1) a *sensor*, 2) an *ObservableProperty*, and 3) an *FOI* for detailing the property with which it was linked.

The knowledge graph provides the benefits of categorizing datasets using multiple observations by defining their metadata fields using state-of-the-art ontologies and vocabularies, and enriching observation concepts using supplementary descriptions using the DCAT vocabulary. The metadata added to observations will extend the technical descriptions of datasets, ultimately improving the search for different datasets and their reuse, interpretation, and integration. Another advantage of using mash-up ontologies for describing datasets can be envisioned: the mapped concepts can help avoid redundancy of collected sensor data. A sensor network comprising different sensors deployed at an FOI may generate thousands of observations. Mapping each acquired observation with its relevant context will increase the number of links and make the requested knowledge graph very large. To avoid this redundancy, the *ObservationCollection* class from SOSA extension ontology is used in our knowledge graph. Constructing collections based on similar environmental observations will reduce the size of the knowledge base and contribute towards the optimization of data storage (i.e. triplestore). However, at present, the functionality of organizing observations using different *ObservationCollections* is not yet implemented while executing the knowledge graph. Future works need to be done to implement the functionality of *ObservationCollections*. Also, the presentation of the subgraphs (generated by Module 3 in the previous section) to the user for studying their metadata needs to be improved. This process will add more user-friendliness to the system's graphical user interface (see Fig. 4), which is not present at the moment. Another limitation of this work is, at present only AGROVOC controlled vocabulary is used by the *PhraseMatcher* of *Spacy* library to categorize the concepts (see Fig. 3) into required metadata fields (e.g. *features*, *methods (procedures)*, *properties* and *systems (sensors))* that exist in *dataset title* and *description*. Advanced text processing functionalities coupled with existing deep learning techniques can be explored for training the *PhraseMatcher* to categorize a broad range of concepts by incorporating multiple controlled vocabularies that are relevant to environmental and life sciences contexts.

4. Conclusion

Numerous datasets are generated daily for research across different scientific disciplines. However, these datasets are still scattered, and are problematic to reuse and integrate. It is a significant challenge to retrieve these datasets, and make it possible to reuse them. Achieving this would promote scientific discovery. To this end, this work has identified the fundamental requirements from a literature review for constructing a dataset discovery knowledge graph for multidisciplinary studies. These requirements were later translated into a set of 9 competency questions. Relevant ontologies and controlled data vocabularies were chosen from the existing literature for creating these questions and for defining the concepts which were used in them. Later, a knowledge graph was instantiated and implemented for environmental and life sciences contexts, aimed at addressing the identified competency questions. An application of this knowledge graph is described using an oceanography example. The major advantage of the developed knowledge graph is that it enables us to retrieve the same dataset using different viewpoints (transitive concepts linked to an actual *FOI*). Also, the knowledge graph has tried to improve the process of dataset discovery using *FOIs*, *ObservableProperty* and *Spatio-temporal representation* irrespective of the scientific domain.

Acknowledgements

The authors would like to acknowledge the Institut de Recherche pour le Développement (IRD) (https://www.ird.fr) for funding this project, the French organization Ifremer for the datasets used for the oceanography use-case, Romain Bouvier for transferring the foundation of this work, Syed Mehtab Alam for his technical advice on Python libraries and Victoria Agazzi for her feedback to this research work.

References

[1] Cornell, S.E., Downy, C.J., Fraser, E.D.G. and Boyd, E. (2009) “Earth system science and society.” Cambridge University Press, 1-38 doi: https://doi.org/10.1017/CBO9780511921155.004.

[2] Steffen, W., Richardson, K., Rockström, J., Schellnhuber, H.J., Dube, O.P., Dutreuil, S., Lenton, T.M., and Lubchenco, J. (2020) “The emergence and evolution of Earth System Science.” *Nature Reviews Earth & Environment* 1(1): 54-63.

[3] O'Riordan, T. (2014) “Environmental science for environmental management.” 2nd edn. Routledge.

[4] Janowicz, K., Haller, A., Cox, S.J., Le Phuoc, D., and Lefrançois, M. (2019) “SOSA: A lightweight ontology for sensors, observations, samples, and actuators.” *Journal of Web Semantics* 56:1-10.

[5] Baumann, K., Bertino, A., Rettig, L., Sigloch, S., Subotic, D., & Subotic, I. “The RESCS Ontology: linking Open Research Data from multiple sources to support interdisciplinary investigations”, available online: http://www.semantic-web-journal.net/system/files/swj2746.pdf

[6] Rayback, S.A., Duncan, J.A., Schaberg, P.G., Kosiba, A.M., Hansen, C.F. and Murakami, P.F. (2020) “The DendroEcological Network: A cyberinfrastructure for the storage, discovery and sharing of tree-ring and associated ecological data.” *Dendrochronologia* 60: 125678.

[7] Riddick, A.T., Heaven, R., Royse, K.R., Singh, A. and Hughes, A.G. (2020) “A model metadata schema for environmental hazard models and its implementation in the PURE portal." *Environmental Modelling & Software* 124: 104597.

[8] Etuk, A., Shaw, F., Gonzalez-Beltran, A., Johnson, D., Laporte, M.A., Rocca-Serra, P., Arnaud, E., Devare, M., Kersey, P.J., Sansone, S.A. and Davey, R.P. (2020), “COPO: A metadata platform for brokering FAIR data in the life sciences.” *BioRxiv* 782771.

[9] Kawashima, S., Katayama, T., Hatanaka, H., Kushida, T., and Takagi, T. (2018) “NBDC RDF portal: a comprehensive repository for semantic data in life sciences.” *Database* 2018, doi: 10.1093/database/bay123.

[10] Hu, W., Qiu, H., Huang, J., and Dumontier, M. (2017) “BioSearch: A semantic search engine for Bio2RDF.” *Database* 2017.

[11] Martin, P., Magagna, B., Liao, X., and Zhao, Z.: Semantic linking of research Infrastructure metadata. In: Towards Interoperable Research Infrastructures for Environmental and Earth Sciences, pp. 226-246. Springer, Cham (2020).

[12] Can, O., Sezer, E., Bursa, O., and Unalir, M.O. (2017) “Comparing relational and ontological triple stores in healthcare domain.” *Entropy* 19(1): 30.

[13] Viqueira, J.R., Villarroya, S., Mera, D., and Taboada, J.A. (2020) “Smart environmental data infrastructures: Bridging the gap between earth sciences and citizens.” *Applied Sciences* 10(3): 856.

[14] Vassallo, V., and Felicetti, A. (2020) “Towards an ontological cross-disciplinary solution for multidisciplinary data: VI-SEEM data management and the FAIR principles.” *International Journal on Digital Libraries*: 1-11.

[15] Williams, Stuart, and Rosecký (2019) “Sensor data and metadata standards review for UKCEH. Wallingford”, UK Centre for Ecology & Hydrology: 81, available online: http://nora.nerc.ac.uk/id/eprint/526628/

[16] Beretta, V., Desconnets, J.C., Mougenot, I., Arslan, M., Barde, J., and Chaffard, V. (2021) “A user-centric metadata model to foster sharing and reuse of multidisciplinary datasets in environmental and life sciences.” *Computers and Geosciences*: 104807

[17] Cox, S. (2020) “Extensions to the Semantic Sensor Network Ontology.” W3C Working Draft, available online: https://www.w3.org/TR/vocab-ssn-ext/, last accessed 2021/04/12.

[18] Leadbetter, A.M., and Vodden, P.N. (2016) “Semantic linking of complex properties, monitoring processes and facilities in web-based representations of the environment.” *International Journal of Digital Earth* 9(3): 300-324.

[19] Maali, F., J. Erickson, and P. Archer. (2020) “Data catalog vocabulary (dcat) W3C recommendation.”, available online: https://www.w3.org/TR/vocab-dcat-2/.

[20] DCMI Usage Board. DCMI Metadata Terms, available online: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

[21] SWEET, Semantic Web for Earth and Environment Technology Ontology, available online: https://bioportal.bioontology.org/ontologies/SWEET.

[22] Cox, S. and Little, C.: Time Ontology in OWL, available online: https://www.w3.org/TR/owl-time/.

[23] Brickley, D. and Guha, R.V.: RDF Schema 1.1, available online: https://www.w3.org/TR/rdf-schema/.

[24] Miles, A., Matthews, B., Wilson, M., and Brickley, D. (2005) “*SKOS core: simple knowledge organisation for the web*.” In: International Conference on Dublin Core and Metadata Applications pp. 3-10.

[25] REPHY – French Observation and Monitoring program for Phytoplankton and Hydrology in coastal waters. REPHY dataset - French Observation and Monitoring program for Phytoplankton and Hydrology in coastal waters. Metropolitan data. SEANOE (2021).

[26] Richardson, Leonard (2007) "Beautiful soup documentation" available online: https://www. crummy. com/software/BeautifulSoup/bs4/doc/.

[27] Reitz, K., I. Cordasco, and N. Prewitt. "Requests: HTTP for humans." https://2. python-requests. org/en/master (2014)

[28] Vasiliev, Y. Natural Language Processing with Python and SpaCy: A Practical Introduction. No Starch Press (2020).

[29] RDFlib, https://github.com/RDFLib/rdflib (2021).

[30] Mehtab Alam, S., E. Arsevska, M. Roche, and M. Teisseire. (2022) “A data-driven score model to assess online news articles in Event-based surveillance system.” In: International Conference on Information Management and Big Data (SIMBig) 2021 (Accepted paper).