The GeoLink Modular Oceanography Ontology

Adila Krisnadhi[1,8], Yingjie Hu[2], Krzysztof Janowicz[2], Pascal Hitzler[1], Robert Arko[3], Suzanne Carbotte[3], Cynthia Chandler[4], Michelle Cheatham[1], Douglas Fils[5], Timothy Finin[6], Peng Ji[3], Matthew Jones[2], Nazifa Karima[1], Kerstin Lehnert[3], Audrey Mickle[4], Thomas Narock[7], Margaret O'Brien[2], Lisa Raymond[4], Adam Shepherd[4], Mark Schildhauer[2], and Peter Wiebe[4]

[1] Wright State University
[2] University of California, Santa Barbara
[3] Lamont-Doherty Earth Observatory, Columbia University
[4] Woods Hole Oceanographic Institution
[5] Consortium for Ocean Leadership
[6] University of Maryland, Baltimore County
[7] Marymount University
[8] Faculty of Computer Science, Universitas Indonesia
krisnadhi.2@wright.edu

Abstract. GeoLink is one of the building block projects within EarthCube, a major effort of the National Science Foundation to establish a next-generation knowledge infrastructure for geosciences. As part of this effort, GeoLink aims to improve data retrieval, reuse, and integration of seven geoscience data repositories through the use of ontologies. In this paper, we report on the GeoLink modular ontology, which consists of an interlinked collection of ontology design patterns engineered as the result of a collaborative modeling effort. We explain our design choices, present selected modeling details, and discuss how data integration can be achieved using the patterns while respecting the existing heterogeneity within the participating repositories.

1 Introduction

Like in other branches of science, data holds a very prominent role in conducting research inquiries in ocean science. A number of synthesis centers sponsored by the National Science Foundation (NSF), such as NCEAS and NESCent, have provided evidences that coupling existing data with interdisciplinary collaboration and analyses can lead to exciting and novel scientific insights, which would be almost impossible to achieve traditionally [3, 5, 11]. This leads to the establishing of ocean (and generally, geo-)science data repositories, such as BCO-DMO, DataONE, and IODP, which contributes to a significant improvement particularly in data preservation. Such data repositories are typically designed to serve specific parts of the geoscience research community, making data management and quality control more tractable. On the flip side, however, data become highly heterogeneus because of the differences in data formats, methods of access, and nuances in the conceptualization. This can cause frustration for researchers when

attempting to find and integrate relevant data from these multiple repositories to perform an integrative analysis [12]. This problem and other related knowledge management problems led to the launching of the EarthCube program by NSF. EarthCube is a major, community-led effort to upgrade cyberinfrastructure for the geosciences consisting of various building block projects and research coordination networks, all aiming to enable extensive cross-discipline data sharing and integration, to allow global data discovery, and to transform the way researchers understand the Earth system via data-enabled geosciences research.

GeoLink[1] is one of the EarthCube building block projects aiming to leverage advances in semantic technologies for developing a data integration and discovery framework involving seven major data repositories, mainly in the area of ocean science. Those repositories are BCO-DMO, DataONE, IEDA, IODP, LTER, MBLWHOI Library, and R2R.[2] The data integration problem faced by this project is both technically and socially challenging, not just because of the lack of direct alignment between data from different repositories, but also due to fundamental differences in the way data and knowledge are modeled. GeoLink tackles this problem by the use of Linked Data [1] and Ontology Design Patterns (ODPs) [2]. Linked Data enables repositories to describe and publish their data using standard syntax featuring links to other data, possibly in different repositories. Meanwhile, ODPs allows a horizontal integration featuring semantic alignment between repositories with possibly independent semantic models.

In this paper, we present the GeoLink modular ontology, which is actually a collection of ODPs developed for the purpose of data integration in the GeoLink project. Before describing the ontology, we start by explaining key points in our modeling approach using ODPs in section 2. Sections 3 and 4 present the ontology and selected modeling details. Due to space restriction, we cannot present the whole ontology in detail, and refer the reader to the more detailed technical report at `http://schema.geolink.org/`. Section 5 describes the availability and external links of this ontology. Finally, Section 6 summarizes this work.

2 Modeling Approach

The GeoLink project aims to provide a framework for horizontal integration amongst data providers. The project, however, does not advocate the creation of an overarching upper ontology for the ocean science because fundamental differences in data modeling and vocabularies between repositories due to differing subdomains, purposes and requirements prevent the realization of such an ontology. Instead, we set out with developing *ontology design patterns* (*ODPs*), or more specifically the so-called *content patterns*, each of which is a self-contained, highly modular ontology encapsulating a particular notion within some domain of discourse and can act as a building block of a more complex ontology [2].

The modeling task was conducted through collaborative modeling sessions, which ensure a good *community engagement* by a very active involvement of

[1] `http://www.geolink.org/`
[2] See `http://www.geolink.org/team.html`.

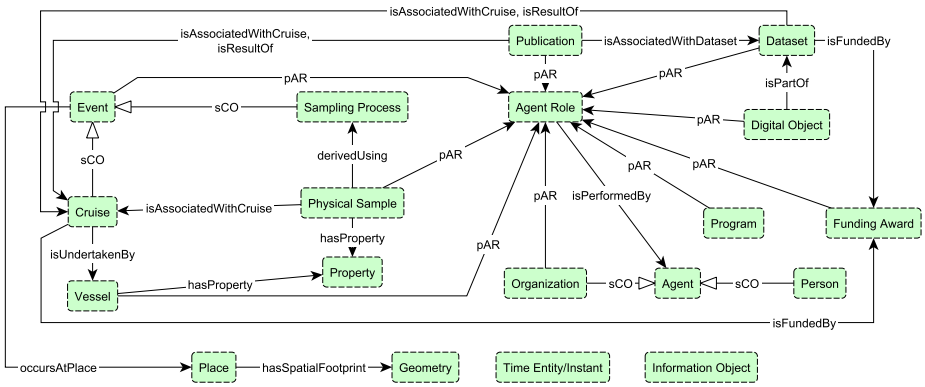

Fig. 1. Schema diagram containing (almost) all patterns in the GeoLink ontology and their main links. All patterns have links to Time Entity/Instant and Information Object, but they are not displayed to make the figure less clutterred. sCO=subClassOf; pAR=providesAgentRole; each box is a pattern, represented by its main class.

oceanographers as domain experts and potential end users. The modeling sessions are intended to bridge language and perspective gaps between ontology engineers, domain experts, and end users, which are bound to occur during an ontology development [7]. In a modeling session, we proceeded by focusing on one notion a time, starting from (i) gathering use cases through a set of competency questions [10]; (ii) identifying and visualizing relevant classes and relationships while keeping within the boundary of the focus notion; and (iii) specifying constraints and axioms, initially in semi-formal natural language expressions. Ontology engineers then continued the work by translating the modeling result from the steps above into a formal ontology, while ensuring no axiom makes an overly strong ontological commitment. Since modeling was focused on one notion at a time, we obtained self-contained, highly modular ontology patterns.

3 Ontology Overview

The GeoLink modular ontology comprises of several content patterns (Figure 1). The majority of the content patterns model some concrete notion deemed important by the participating data providers as it reflects an important discovery facet. These include cruise, person, organization, dataset, funding award, program, etc. A few other content patterns represent some form of abstraction introduced typically as a good modeling practice or as a flexible connector between two other patterns. The remaining patterns are auxiliary content patterns that provide more details to some other content patterns. We briefly present an overview of each of these patterns in the following and the reader is referred to the technical report and our OWL implementation for more details.

We start with the abstract patterns **Agent** and **Agent Role**. The Agent pattern defines a central class Agent and allows one to express that an agent

(e.g., a person or an organization) may perform a role, which is an instance of the AgentRole class. The latter is aligned to the AgentRole class in the Agent Role pattern. The Agent Role pattern itself is essentially a reification of relations between an agent and the thing the agent is involved in. For example, a person may participate in a cruise as a chief scientist. Then, using the Agent Role pattern, we express this by stating that a cruise provides a role of type chief scientist that is performed by that person. This reification allows to flexibly cover various ways in which an agent may be related to the thing the agent is involved in. The pattern also allows us to express the starting and ending time of a role.

The abstract pattern **Event** describes generic events, which may include cruises, sampling processes, etc. This pattern is inspired by the Simple Event Model (SEM) [4], but augmented with a stronger axiomatization in OWL. In this model, an event is something that occurs at some place and some time, and may provide agent roles performed by agents.

The abstract pattern **Information Object**, reused from DOLCE [8], encapsulates information commonly attributed to an object, including name, aliases, description, webpage, and other non-URI identifiers. This pattern allows us to collect many pieces of information about an object that become relevant when it is understood as an information artifact.

The **Place** pattern captures spatial information in the Event pattern above and the rest of the ontology. It expresses that a place has a geometry as its spatial footprint, similar to the relationship between geographic feature and geometry in GeoSPARQL [9]. The Information Object pattern is used to represent other information about a place such as its name and description.

The **Person** pattern is a specialization of the Agent pattern, describing human persons. As an Agent, a person may perform a role in a particular context. Additionally, the Person pattern allows one to say that a person has personal information items. A personal information item is an attribute of a person, such as name, address, etc., that may change during his/her lifetime, and is modeled through the auxiliary **Personal Info Item** pattern. The **Person Name** pattern is also defined as a specialization of the Personal Info Item pattern.

The **Organization** pattern is a specialization of the Agent pattern, describing organizations, including academic institutions, funding agencies, vessel owners, etc. This pattern also models affiliation relationships of an agent to organizations using the Agent Role pattern. Every organization is described by exactly one information object that encapsulates additional information about the organization such as name and location. This last part is modeled by reusing and aligning with the Information Object pattern.

The **Cruise** pattern describes oceanographic cruises. A cruise is modeled as a type of event whose spatiotemporal component is determined by its trajectory. The Agent Role pattern is used to model various roles a person or organization may hold in relation to a cruise. In addition, the **Vessel** pattern models vessels, which is the physical object with which the cruise is undertaken.

The **Funding Award** pattern describes funding awards given to researchers to carry out their ocean science research activities. It has a starting and ending date and provides roles to agents such as principal investigator, sponsor, etc.

The **Program** pattern captures the notion of ocean science programs. A program is a loose group of activities, funding awards, and other things related to ocean science research that follow certain scientific themes or objectives. It can be in the form of a strategic initiative spanning different projects, or a collaborative network involving many scientists working on different projects which share some common strategic goals.

The **Dataset** pattern models the notion of dataset and common metadata such as description, creator, creation time, etc. A dataset can be associated with features of interests and may contain digital objects. The **Digital Object** pattern represents digital objects, which are understood as file-like objects within a data repository.

The **Physical Sample** pattern minimally represents discrete specimens (rocks, sediments, fluids, etc) collected from the natural environment for scientific study.

4 Selected Modeling Details

In this section, we present selected modeling details from the Cruise pattern, which is arguably one of the most interesting parts of the GeoLink ontology. Oceanographic cruises indeed play a very central role in the professional lives of many ocean scientists, and so it is very natural to utilize them as an aspect of data organization and sharing. In our ontology, an oceanographic cruise is modeled as a type of event whose spatiotemporal component is represented by its trajectory. Like in the Event pattern (Fig. 2), the Cruise pattern employs the Agent Role pattern to model involvement of agents in it (Fig. 3). Alignment of Cruise pattern to the Event pattern is specified through axioms written in description logic (DL) notation in (1)-(5) with glev: denoting the namespace prefix of the Event pattern. Specific roles for cruise are defined by subclassing AgentRole.

$$\mathsf{Cruise} \sqsubseteq \mathsf{glev{:}Event},\ \mathsf{Port} \sqsubseteq \mathsf{glev{:}Place},\ \mathsf{TimeEntity} \sqsubseteq \mathsf{glev{:}TimeEntity} \quad (1)$$

$$\mathsf{AgentRole} \sqsubseteq \mathsf{glev{:}AgentRole},\ \mathsf{Agent} \sqsubseteq \mathsf{glev{:}Agent} \quad (2)$$

$$\mathsf{hasTrajectory} \circ \mathsf{hasFix} \circ \mathsf{hasLocation} \circ \mathsf{hasSpatialFootprint}^{-} \sqsubseteq \mathsf{glev{:}occursAtPlace} \quad (3)$$

$$\mathsf{hasTrajectory} \circ \mathsf{hasFix} \circ \mathsf{atTime} \sqsubseteq \mathsf{glev{:}occursAtTime} \quad (4)$$

$$\mathsf{providesAgentRole} \sqsubseteq \mathsf{glev{:}providesAgentRole},\ \mathsf{isPerformedBy} \sqsubseteq \mathsf{glev{:}isPerformedBy} \quad (5)$$

We then assert that a cruise has exactly one trajectory, is undertaken by exactly one vessel, and is described by exactly one InformationObject. Additionally, this instance of InformationObject describes exactly only the cruise. This InformationObject, which is aligned to the class of the same name from the Information Object pattern, acts as a proxy through which we describe various information about the cruise not covered by having the cruise as an event. In addition to the data properties (not displayed in the figure) inherited from the

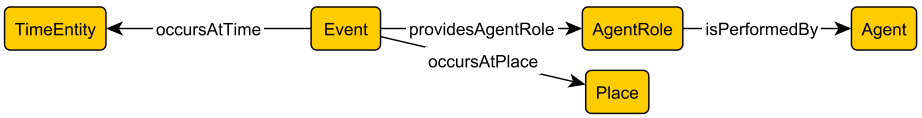

Fig. 2. Event pattern

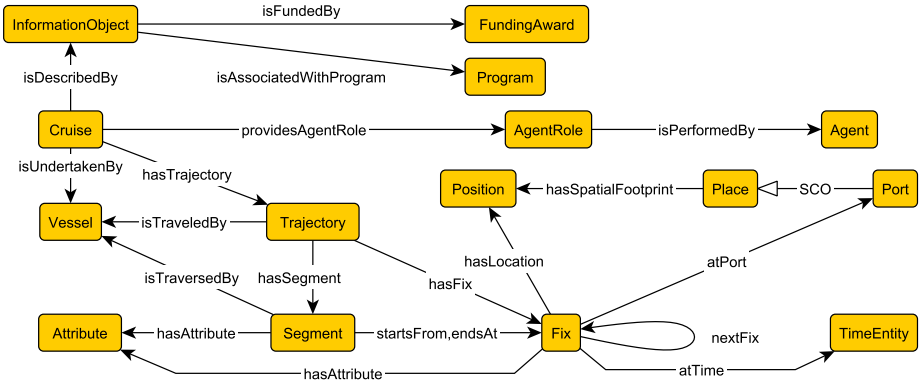

Fig. 3. Cruise with Trajectory and Agent Roles. SCO=subClassOf

Information Object pattern such as identifier, description, webpage, etc., an InformationObject of a cruise may carry information regarding funding award, program, as well as other cruises related to the cruise described by this instance of InformationObject. We also assert that if a cruise is undertaken by a vessel, then this vessel has to travel along the cruise's trajectory.

$$\mathsf{Cruise} \sqsubseteq (=1\ \mathsf{hasTrajectory}.\mathsf{Trajectory}) \sqcap (=1\ \mathsf{isUndertakenBy}.\mathsf{Vessel}) \sqcap (=1\ \mathsf{isDescribedBy}.\mathsf{InformationObject}) \tag{6}$$

$$\mathsf{InformationObject} \sqsubseteq (=1\ \mathsf{isDescribedBy}^{-}.\mathsf{Cruise}) \tag{7}$$

$$\mathsf{hasTrajectory}^{-} \circ \mathsf{isUndertakenBy} \sqsubseteq \mathsf{isTraveledBy} \tag{8}$$

A cruise trajectory in turn represents a route that the cruise takes. To model the notion of cruise trajectory, we reused and extended the Semantic Trajectory pattern, which already provides basic vocabulary and OWL axiomatization [6]. Generally, a trajectory is given by an ordered collection of *fixes*, representing time-stamped locations. Non-spatiotemporal information specific to a fix can be included via its *attributes*, for example, to indicate that the fix is the arrival to some port stop. Between two consecutive fixes, we define a *segment*, traversed by some moving object, e.g., a vessel. Details of the axiomatization for trajectory is given in the technical report and in the OWL implementation.

One other piece of detail we would like to convey is the use of guarded domain and range restrictions. Most of the arrows in Figures 2 and 3 represent object properties and the direction of the arrows goes from the domain part of the property towards its range. A straightforward way to axiomatize

these would have been as unguarded domain and range restrctions, i.e., as axioms of the form P rdfs:domain C and P rdfs:range C, which are equivalent to $\exists P.\top \sqsubseteq C$ and $\top \sqsubseteq \forall P.C$ using DL notation. For patterns, however, this would introduce rather strong ontological commitments which would make future reuse of the patterns more difficult. Hence, we use guarded versions of the restrictions, e.g., for the hasTrajectory property, we state the two axioms $\exists \mathsf{hasTrajectory}.\mathsf{Trajectory} \sqsubseteq \mathsf{Cruise}$ and $\mathsf{Cruise} \sqsubseteq \forall \mathsf{hasTrajectory}.\mathsf{Trajectory}$.

5 Availability

All ODPs for GeoLink are available online from `http://schema.geolink.org/`, including the technical report containing detailed descriptions of all patterns. This will be made available even beyond the duration of the GeoLink project because the participating repositories have committed to continue the integration effort, which is also strongly motivated by ocean science researchers who have been using these repositories for their research activities.

Each pattern resides in its own OWL file with ontology URI of the form `http://schema.geolink.org/version/pattern/[patternname]`. For instance, `http://schema.geolink.org/dev/pattern/agentrole` is the URI where the Agent Role pattern currently resides. The `dev` part in the URI indicates that the pattern is currently under development. A stable release will replace `dev` with a version number. A pattern is typically aligned to another pattern or an external ontology, and this is incorporated by creating a separate OWL file containing axioms for one direction of alignment. For example, the alignment from the Cruise pattern to the Event pattern is provided by the module at `http://schema.geolink.org/dev/pattern/cruise-to-event`, which really contains axioms (1)-(5). Meanwhile, an alignment to the W3C Time ontology[3] is provided by the module at `http://schema.geolink.org/dev/pattern/cruise-to-owltime`.

Note that such alignment modules specify one-direction alignment as they do not specify an alignment from the opposite direction. For example, the module at `http://schema.geolink.org/dev/pattern/cruise-to-event` really only specifies an alignment from Cruise to Event, and not from Event to Cruise. In terms of file organization, such alignment modules are imported by the pattern that is the origin of the alignment. So, the module at `http://schema.geolink.org/dev/pattern/cruise-to-event` is imported by the Cruise pattern at `http://schema.geolink.org/dev/pattern/cruise`.

Besides the W3C Time ontology, external linkages also exist to other ontologies and vocabularies. The Place pattern is aligned to the GeoSPARQL ontology,[4] in particular the Geometry class. Standard geographic features from the GEBCO gazetteers[5] are used to populate the Place pattern. In addition, we adopt parts of the SeaVoX standard platform types[6] to obtain types of vessels.

[3] `http://www.w3.org/2006/time`
[4] `http://www.opengis.net/ont/geosparql`
[5] `http://www.gebco.net/data_and_products/undersea_feature_names/`
[6] `http://vocab.nerc.ac.uk/collection/L06/current/`

6 Conclusions

We have presented the GeoLink ontology, consisting of a number of ODPs designed for the oceanography domain. The resulting patterns are sufficiently modular, and thus arguably easier to extend than foundational, top-level ontologies. Currently, the GeoLink project is in the middle of populating the patterns with real data and a very preliminary evaluation demonstrated that the patterns together can serve as an integrating layer of heterogeneous ocenographic data repositories.

Acknowledgement. The presented work has been primarily funded by the National Science Foundation under the award 1440202 "EarthCube Building Blocks: Collaborative Proposal: GeoLink – Leveraging Semantics and Linked Data for Data Sharing and Discovery in the Geosciences."

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)
2. Gangemi, A.: Ontology design patterns for semantic web content. In: Gil, Y., et al. (eds.) The Semantic Web - ISWC 2005, 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, November 6-10, 2005, Proceedings. Lecture Notes in Computer Science, vol. 3729, pp. 262–276. Springer (2005)
3. Hackett, E.J., et al.: Ecology transformed: NCEAS and changing patterns of ecological research. In: Olson, G.M., et al. (eds.) Science on the Internet, pp. 277–296. MIT Press, Cambridge, MA (2008)
4. van Hage, W.R., et al.: Design and use of the Simple Event Model (SEM). Journal of Web Semantics 9(2), 128–136 (2011)
5. Hampton, S.E., Parker, J.N.: Collaboration and productivity in scientific synthesis. BioScience 61(11), 900–910 (2011)
6. Hu, Y., et al.: A geo-ontology design pattern for semantic trajectories. In: Tenbrink, T., et al. (eds.) Spatial Information Theory – 11th International Conference, COSIT 2013, Scarborough, UK, September 2-6, 2013. Proceedings. Lecture Notes in Computer Science, vol. 8116, pp. 438–456. Springer, Heidelberg (2013)
7. Kalbasi, R., et al.: Collaborative ontology development for the geosciences. Transactions in GIS 18(6), 834–851 (2014)
8. Oberle, D., et al.: DOLCE ergo SUMO: On foundational and domain models in the SmartWeb Integrated Ontology (SWIntO). Journal of Web Semantics 5(3), 156–174 (2007)
9. Open Geospatial Consortium: OGC GeoSPARQL - a geographic query language for RDF data. Open Geospatial Consortium (2011), Document 11-052r4
10. Presutti, V., et al.: eXtreme Design with content ontology design patterns. In: Blomqvist, E., et al. (eds.) Proceedings of the Workshop on Ontology Patterns (WOP 2009), Washington D.C., USA, 25 October 2009. CEUR Workshop Proceedings, vol. 516. CEUR-WS.org (2009)
11. Reichman, O., Jones, M.B., Schildhauer, M.P.: Challenges and opportunities of open data in ecology. Science 331(6018) (2011)
12. You, J.: Geoscientists aim to magnify specialized web searching. Science 347(6217), 11–11 (2015)