The GeoLink Framework for Pattern-based Linked Data Integration

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

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. Specifically, GeoLink aims to improve data reuse and integration of seven geoscience data repositories through the use of ontologies. In this paper, we present the approach taken by this project, which combines linked data publishing and modular ontology engineering based on ontology design patterns to realize integration while respecting existing heterogeneity within the participating repositories.

1 Introduction

With the establishment dozens of data repositories, data integration is becoming a major challenge faced by the ocean (and geo-)science research community. The problem stemmed from the fact that data repositories were established to serve specific parts of the community, which leads to a very high degree of data heterogeneity in formats, methods of access, and conceptualization. GeoLink project[1], a part of EarthCube, the National Science Foundation (NSF)'s larger effort to establish next-generation knowledge infrastructure, aims to develop a flexible and extendible data integration framework, starting from seven major ocean science data repositories[2] by leveraging Linked Data [1] and Ontology Design Patterns (ODPs) [2]. With Linked Data, repositories describe and publish their data using standard model that includes links to other data in other repositories. Meanwhile, horizontal alignment across different repositories with possibly

[1] See `www.geolink.org`, `schema.geolink.org`, and `data.geolink.org`

[2] BCO-DMO, DataONE, IEDA, IODP, LTER, MBLWHOI Library, and R2R

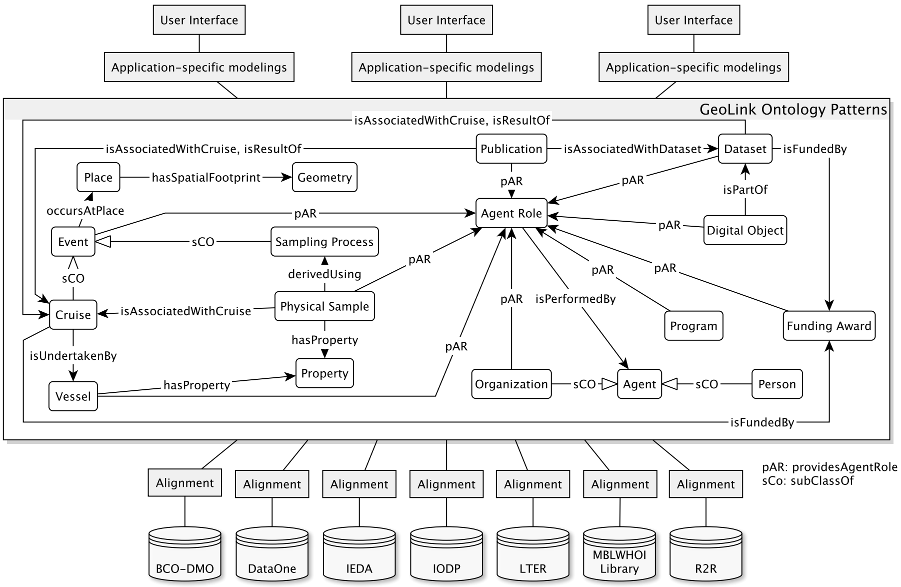

Fig. 1: GeoLink data integration framework with some details on the ODP layer.

independent semantic models can be achieved with the help of ODPs. In this paper, we present an ODP-based data integration employed in GeoLink and we specifically invite the conference attendees to our corresponding poster presentation since this paper complements our ISWC 2015 ontology paper [3] – which describes the details of the GeoLink ODP collection – with the description of GeoLink architecture for cross-repository discovery that still provides sufficient flexibility and extendibility for data providers.

2 Data Integration Framework with ODPs

The GeoLink data integration framework, depicted in Figure 1, has essentially three layers: the data sources/repositories, the global schema, and the user interface. Here, the data sources are assumed to be *linked data repositories*, hence at least, we have RDF as the standard data model. The more challenging problem is integration at the level of semantic. Intuitively, the global schema's role is to provide a common vocabulary that a user can use to perform data discovery.

From the outset, this three-layered framework is similar to your typical ontology-based data integration framework. In such a framework, the global schema is realized in the form of *one* ontology, typically a monolithic, upper-level ontology. The problem with this typical approach is that such an ontology is too cumbersome to use and very hard to understand, especially for data providers who may not have the necessary expertise. Moreover, semantic heterogeneity in the data across different repositories exponentially increases the difficulty in using and maintaining the ontology. When a new data repository wishes to join

```
CONSTRUCT {
  ?x a :Cruise ;
     :providesAgentRole [a :ChiefScientistRole; :isPerformedBy ?p ] .
} WHERE {
  ?x a bcdmo:Deployment ;  bcdmo:ofPlatform [a bcdmo:Vessel] ;
     bcodmo:hasChiefScientist ?p . }
```

Fig. 2: Example query for populating ODPs

the framework, a complicated adjustment of the ontology may become necessary to ensure the existing integration does not break.

To alleviate this problem, GeoLink framework employs a *set* of *ODPs* as the global schema, instead of one monolitihic ontology. Each ODP models a generic notion in a particular domain. So, the key part of the approach is identifying a number of generic notions in ocean science relevant for the data repositories involved in the project. The project then proceeded by modeling each notion one by one collaboratively in a modular way, identifying widely reusable and reoccurring aspects in those notions. Each notion gives us one ODP realized as a self-contained, highly modular ontology, which are sufficient to define the given notion precisely without putting too strong ontological commitments. A high-level overview of the set of ODPs for GeoLink can be seen in the middle layer of Figure 1. Details of these ODPs are given in [3], including collaborative modeling efforts between the ontology engineers and the domain experts needed to make sure the ODPs are *grounded* in representative, real use cases.

Data providers can then flexibly join this integration framework by *populating the ODPs*: making RDF triples representing their data are annotated with vocabulary from ODPs and available as linked data. This can be done in different ways, e.g., exposing the data via SPARQL endpoint or providing dumps of RDF triples. This constitutes the intermediate layer of "alignment" between the data repositories and the ODPs in Figure 1. From users' perspective, if the data from all data providers are annotated with the ODPs, they will in principle only see one RDF graph (not necessarily residing in a central hub), which aggregates data from all participating repositories. Vocabulary in the ODPs can then be used as the language with which *federated queries* to the data can be formulated. For data providers, however, populating ODPs may not be as straightforward.

Essentially, the GeoLink framework offers two approaches for data providers to populate the ODPs. ***First approach***, data providers annotate their data by *directly* employing the vocabulary defined by the ODPs. Specification of the vocabulary can be easily obtained as OWL files from GeoLink. Unfortunately, it is possible that some data providers, especially if they join already have their own linked data schema or use their own vocabulary of choice, are reluctant to do the first approach. So, the ***second approach*** is that data providers provide a *schema mapping* to GeoLink ODPs. Such a schema mapping can be expressed using a SPARQL CONSTRUCT query, which can either be used to generate RDF triples in batch or on the fly. For example, the schema in BCO-DMO does not contain the class Cruise explicitly, but rather, understands Cruise as a Deployment whose platform is a Vessel. The Cruise class in the Cruise ODP can thus be populated by executing a query like the one in Figure 2. The query also generates a new node for ChiefScientistRole since BCO-DMO models chief scientist of a Deployment using a property hasChiefScientist. This query illustrates the flexbility of the framework since data providers do not need to change their schema, nor the patterns need to be modified so long as such a data transformation query can be written. This second approach also opens up the possibility for data providers who prefer repository-specific schema over direct use of the ODPs, hence lowering the barrier of integration.

3 Conclusions and Outlook

We have presented a data integration framework based on ontology design patterns. The use of ODPs allows us to achieve cross-repository discovery, while respecting semantic heterogeneity residing within each repository. For future work, in the context of GeoLink project, we plan to reach out to more partners from other EarthCube projects to participate in the framework, also to test the effectiveness and robustness of the approach, which may include extending the current set of ODPs. We also plan to explore possibilities to automate some parts of the framework, for instance, leveraging advances in ontology alignment to help data providers establish alignment to the patterns. We are also looking at different computational issues with the implementation of the framework as well as bringing reasoning into the picture, e.g., for detecting inconsistency and incompleteness in the data, or smarter discovery. Finally, we also plan to do a usability test from the perspective of the data consumers, i.e., the geoscientists.

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. Krisnadhi, A., Hu, Y., Janowicz, K., Hitzler, P., Arko, R., Carbotte, S., Chandler, C., Cheatham, M., Fils, D., Finin, T., Ji, P., Jones, M., Karima, N., Lehnert, K., Mickle, A., Narock, T., O'Brien, M., Raymond, L., Shepherd, A., Schildhauer, M., Wiebe, P.: The GeoLink modular oceanography ontology. In: The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference, Betlehem, PA, USA, October 11-15, 2015 (2015), accepted for publication.