DATA ARTICLE

The GeoLink Knowledge Graph

Michelle Cheatham[a], Adila Krisnadhi[b], Reihaneh Amini[a], Pascal Hitzler[a], Krzysztof Janowicz[c], Adam Shepherd[d], Tom Narock[e], Matt Jones[f], Peng Ji[g]

[a]Wright State University, Dayton, OH, USA; [b]Universitas Indonesia, Depok, Indonesia; [c]University of California at Santa Barbara, Santa Barbara, CA, USA; [d]Woods Hole Oceanographic Institution, Woods Hole, MA, USA; [e]Notre Dame of Maryland University, Baltimore, MD, USA; [f]National Center for Ecological Analysis and Synthesis, Santa Barbara, CA, USA; [g]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA

ARTICLE HISTORY
Compiled April 20, 2018

ABSTRACT
GeoLink has leveraged linked data principles to create a dataset that allows users to seamlessly query and reason over some of the most prominent geoscience metadata repositories in the United States. The GeoLink dataset includes such diverse information as port calls made by oceanographic cruises, physical sample metadata, research project funding and staffing, and authorship of technical reports. The data has been published according to best practices for linked data and is publicly available via a SPARQL endpoint that at present contains more than 45 million RDF triples together with a collection of ontologies and geo-visualization tools. This paper describes the geoscience datasets, the modeling and publication process, and current uses of the dataset. The focus is on providing enough detail to enable researchers, application developers and others who wish to leverage the GeoLink data in their own work to do so. The dataset is available at http://hdl.handle.net/1912/9524.

KEYWORDS
geoscience data; oceanographic data; knowledge graph; linked data; instance matching; ontology

1. Introduction

Often the most innovative and useful discoveries come at the intersection of traditional fields of research. This is particularly true in the geosciences, which bring together disparate groups of researchers such as geologists, meteorologists, climatologists, ecologists, archeologists, and so on. The National Science Foundation (NSF) has recognized this, and in 2011 it launched the EarthCube initiative.[1] EarthCube is funding multiple interdisciplinary research teams through at least 2022 in an effort to establish a bottom-up approach to creating standards and infrastructure to facilitate collaboration across traditional geoscience domains. One project funded as part of EarthCube is GeoLink.

An enormous amount of data is created every day by many different sources. Previously, this information was available in a wide variety of forms and formats and could not be accessed in a consistent and unified manner. Attempts to address this obstacle

CONTACT Michelle Cheatham. Email: michelle.cheatham@wright.edu

to data integration led to the rise of linked data Berners-Lee (2006). The goal was to establish a set of standards and techniques to make data consumable by both humans and machines in a uniform way. GeoLink has leveraged linked data principles to create a knowledge graph that allows users to seamlessly query and reason over some of the largest geoscience data repositories in the United States. The GeoLink knowledge graph includes such diverse information as port calls made by oceanographic cruises, physical sample metadata, research project funding and staffing, and authorship of technical reports. The graph was already quite large at its inception and has been growing over time. In 2015 it contained over 10 million triples; by 2017, that number had risen to 48 million.

In this paper we present the GeoLink knowledge graph, from the data it contains, to how it is modeled, to how to access it. Our focus is on providing enough detail to enable researchers, application developers and others who wish to leverage this data in their own work to do so. We also highlight existing uses of the knowledge graph and its relationship to other important geoscience vocabularies.

2. GeoLink Data Providers

The datasets that make up the GeoLink knowledge graph represent a huge investment on the part of the geoscience research community. Many of them are funded by large government organizations, such as the NSF. In some cases, NSF funding of individual researchers is contingent upon them publishing their data in one or more of these repositories.

The datasets that currently comprise the GeoLink knowledge graph are:

R2R The Rolling Deck to Repository (R2R)[2] program is the steward of environmental sensor data collected by the U.S. academic research fleet. R2R maintains a catalog of vessels, instrument systems, expeditions, datasets, investigators, organizations, funding awards, cruise reports, and navigation tracks for research cruises.

BCO-DMO The Biological and Chemical Oceanography Data Management Office (BCO-DMO)[3] manages data and information generated during oceanographic research efforts and publishes these data online. Many of the datasets contain measurements collected during research cruises on board U.S. academic vessels, details about which are present in the R2R dataset.

IODP The International Ocean Discovery Program (IODP)[4] and its predecessor programs involve the collection of sediments, rocks, biota, and fluids from beneath the seafloor. This dataset contains information about these samples and the circumstances under which they were collected.

MBLWHOI Marine Biological Laboratory Woods Hole Oceanographic Institution (MBLWHOI) Library[5] maintains a repository of text documents including technical reports, theses, and journal articles covering topics related to marine life and its environment, along with associated metadata.

SESAR The System for Earth Sample Registration (SESAR)[6] is a collection of metadata about natural samples such as rocks specimens, water samples, and sediment cores. It includes information about where, when, and how a sample was collected.

DataONE The Data Observation Network for Earth (DataONE)[7] provides access to data and metadata from a large number of contributors and repositories related to the earth and environmental sciences. The focus of the project is to provide an integrated portal for the discovery of environmental data for researchers, students, and the general public.

AGU-NSF The American Geophysical Union (AGU) has historical data related to its annual conferences as well as NSF funded proposals related to geophysics, including the award information, the project, and the principal and co-principal investigators. There is a large degree of overlap between the people and projects in this dataset and those in the others in this list.

NGDB The National Geochemical Database (NGDB)[8] contains information about the geochemical content of thousands of samples from American stream sediments, soils, and waters, as well as metadata about the time and place the sample was collected.

USAP The United States Antarctic Program (USAP)[9] Data Center is dedicated to preserving and making available the result of NSF-funded research related to Antarctica and the Southern Ocean. It includes data about ice coverage, snowfall totals, and glacial movements in the region. The information included in GeoLink is at the dataset level and includes information such as the curator and funding agency.

3. Modeling and Publication

The overall goal of GeoLink is to integrate geoscience data from many existing repositories into a unified knowledge graph that can be accessed seamlessly. For instance, if the knowledge graph is queried for a particular scientist, results might include research cruises she has been on (from R2R), datasets she has collected (from BCO-DMO), papers she has written (from MBLWHOI), and funding awards she has been granted (from NSF). Similarly, if a rock specimen is queried for, the user can determine who collected it, when, and under what funding award. In order to facilitate this, an underlying schema had to be developed and content providers had to have a mechanism to publish their data as RDF triples that conformed to this schema. This section describes that process. For additional information see Krisnadhi et al. (2015a).

3.1. The GeoLink Modular Ontology

The GeoLink schema rests upon the development of a set of ontology design patterns (ODPs), each of which is a self-contained, highly modular ontology snippet encapsulating a concept, such as person or a physical sample, that occurs within many geoscience repositories. These ODPs were collaboratively developed by a group of ontologists, domain experts, and data providers. The patterns were then stitched together to form the GeoLink modular ontology (GMO). The ODPs represent the concepts within GeoLink that unite the different data repositories, and are therefore the aspects where integrated querying and reasoning are vital to achieving the capabilities laid out at the beginning of this section. Data providers can publish the parts of their data related to these core concepts according to the vocabulary of the GMO. Any elements in a repository that are not related to the GMO are published in the external vocabulary

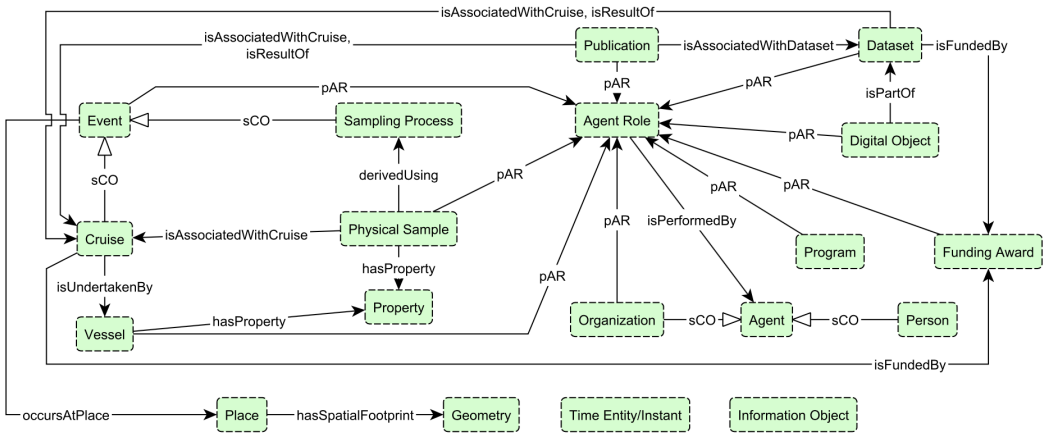

Figure 1. Schema diagram containing (almost) all patterns in the GeoLink ontology and their main links. All patterns have links to Time Entity/Instant and Information Object, but they have been omitted for clarity. sCO=subClassOf; pAR=providesAgentRole; each box is a pattern, represented by its main class.

that best fits the data. This avoids the need for providers to shoehorn their data into a schema that does not fit. The significant patterns within the GMO are shown in Figure 1. The GMO is more fully described in Krisnadhi et al. (2015b).

3.2. The GeoLink Base Ontology

While the GMO follows best-practices in ontological modeling, see Hitzler, Gangemi, Janowicz, Krisnadhi, and Presutti (2016), the data providers found it difficult to work with. For example, the GMO contains an Agent Role pattern (Figure 2) that reifies some relationships. For instance, there is a subclass of AgentRole called SponsorRole, which is *provided by* a funding award and *performed by* an organization. Subclassing AgentRole allows new roles to be added easily in the future. It also enables queries such as what organization that a sponsor worked for *when a project took place*, that would not be possible if this n-ary relationship was represented as a binary one (because the person may have taken a new job after the project ended, for example). The data providers had difficulty applying this schema, because looking at their own datasets, they found nothing equivalent to AgentRole, and looking at the GMO, they found no obvious way to model the Sponsor field in their database. Additionally, reification led to the generation of blank nodes and the need to create and maintain many URIs.

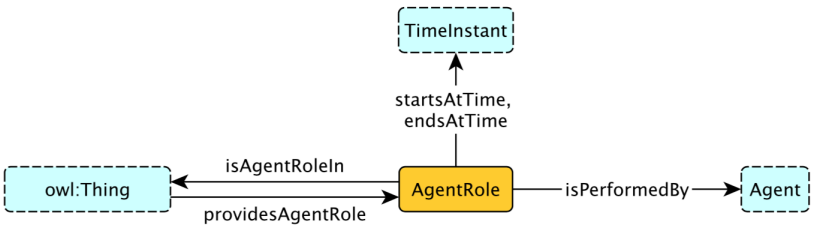

Figure 2. The Agent Role pattern

To handle this issue, a second ontology, termed the GeoLink Base Ontology (GBO), was developed. The GBO is a "flatter", more simplified representation of the GMO. Continuing the AgentRole example from above, the subclasses of AgentRole in the GMO were systematically used to create new properties within the GBO. For instance, the solid lines in Figure 3 show properties related to SponsorRole (a subclass of AgentRole) that are present in the GMO. The dashed line shows a new property, hasSponsor, that is present in the GBO in lieu of the GMO properties.

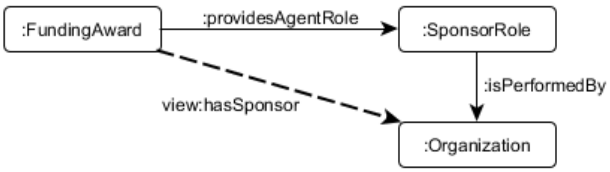

Figure 3. Example of a GBO property derived from the AgentRole pattern

The GBO is less extensible, of less quality regarding good modeling practice, and less suitable to scalable reasoning, but it is easier for the data publishers to work with. Data providers align the relevant parts of their data to the GBO, which is the schema underlying this dataset. A set of SPARQL construct queries can then be used to convert the information into the GMO's schema. Both the GMO and the GBO, along with the alignment between them, can be downloaded from `http://doi.org/10.6084/m9.figshare.5907172` under a CC-BY License.

The GBO leverages several widely-used controlled vocabularies from the geosciences. For example, the GEBCO set of terms are included in the GBO as subclasses of GeoFeature. Additionally, the National Environment Research Council's (NERC's)[10] L05 and L22 vocabularies are used for instruments and instrument types in the GBO, P02 and P03 are used for measurements, and L06 is used for platforms.

As mentioned previously, data providers can use their own schemas for data that is outside the scope of the GBO. Several providers have already done this by leveraging a combination of their own ontologies and existing vocabularies such as FOAF,[11] Dublin Core,[12] and Prov-O.[13] Including this provider-specific information enhances the utility of GeoLink by allowing users to query across multiple datasets to find the resource in which they are interested and then drill down to explore *all* of the details available about that resource.

3.3. Publishing and Harvesting

In order to minimize query response time, the GeoLink data is stored on a centralized server. Individual data providers publish "data dumps" that contain their data, formatted according to the GBO, as RDF triples. These data dumps are collected by a utility called the Harvester and published to the centralized server.

Providers are responsible for making their data dumps available and accessible to the Harvester through an HTTP endpoint, which is described in a VoID configuration file along with the last modified date of the dump. Providers must manually register their VoID file with the Harvester. New dumps can be created at any time. Because most triples stay the same between publishing runs, providers may provide partial dumps with only the triples that have been added since the last harvest.

The Harvester is alerted to any update of a provider's data dump by means of an

updated last modified date in the provider's VoID entry. In the event of an update, the Harvester retrieves the dumps and stores them in a "staging" triple store. It validates them to ensure they are well-formed RDF that comply with the semantics of the GBO. A basic coreference resolution algorithm is then applied to the new triples to produce internal links to existing data. Because the automated coreference resolution is not perfect, these coreferences are not represented using owl:sameAs. Instead, the predicate gl:weakCoreferentOf is used for cases in which the primary label is the same (e.g. name for persons, title for papers, etc.) and gl:strongCoreferentOf indicates that the primary label is the same and at least one additional property matches (e.g. name and email address for persons, title and publication date for papers). Finally, the new triples are promoted to the production triple store.

3.4. Linking

Many of the repositories that make up GeoLink have overlapping content. For example, information about oceanographic cruises occurs in R2R, BCO-DMO, and SESAR, while nearly all datasets have information about researchers, projects, and funding agencies. Linking the same cruise, person, or funding organization across the different repositories is what enables integrated querying and makes GeoLink so useful. Some of these links have been generated manually by the data providers themselves, while others have been created using an automated coreference resolution algorithm as described in Section 3.3. There are thousands of internal links within the GeoLink knowledge graph.

Regarding links from GeoLink to external datasets, right now these take two forms: Open Researcher and Contributor Identifiers (ORCiDs) and Digital Object Identifiers (DOIs). ORCiDs are minted by an open not-for-profit organization at no cost to researchers.[14] They are intended to uniquely identify researchers so that those individuals can be correctly credited for their research work and links can be provided to express their professional affiliations. These IDs are already in use by several major institutions, including all IEEE journals, where they are required, and the NSF, where they are highly encouraged. ORCiDs contained within GeoLink can thus be used to unambiguously identify the corresponding individuals in datasets associated with these other institutions. Similarly, DOIs are persistent identifiers used to provide a stable way to access digital objects whose actual location (e.g. URL) may change over time.[15] These are typically assigned by publishers when a new article, report, or dataset is produced. Because all organizations may refer to a digital resource via its DOI, this can be used to link the same resource across multiple datasets.

4. Discussion of Data Quality

While it is difficult to directly evaluate the quality of a knowledge graph, this section examines the GeoLink knowledge graph from several different viewpoints.

4.1. Five Star Linked Data

Tim Berners-Lee established a five star rating system to evaluate the quality of a linked dataset, see Berners-Lee (2006). The ratings are shown below. To achieve a particular star rating, the dataset must meet the criterion for that rating, plus the criteria for

all lower ratings.

One Star: Web-accessible
Two Star: Structured and machine-readable format
Three Star: Non-proprietary format
Four Star: Based on W3C standards
Five Star: Linked to other datasets to provide context

The GeoLink knowledge graph clearly meets the first four of these criteria: the dataset is available as described in Section 5, it is published as RDF triples, which are structured, machine-readable, non-propriety, and a W3C standard, and the ontology is based on OWL and RDF, also W3C standards. As the previous section shows, the knowledge graph also contains thousands of links, both internal and external, that provide context to the data and raise it to the five star level.

4.2. Coverage

Section 3 began by describing a common use case that GeoLink is intended to support: querying to find *all* of the data about an entity that is available across the constituent data repositories. A web interface (Figure 4) is available at demo.geolink.org that allows a user to do this, as well as to explore the combined linked data from all participating GeoLink repositories by following links between them. While no formal precision and recall metrics are available, due to the lack of a reliable gold standard, use by domain experts both on the team and within the larger geoscience community at NSF and AGU demonstrations has elicited positive feedback regarding the graph's coverage.

4.3. Geoscience Usage

Wider usage of the knowledge graph, beyond the originally targeted use case, further supports the argument about its utility. For instance, a set of web components has been developed and made publicly available[16] to allow web developers to easily incorporate GeoLink data into their site using very basic HTML. For example, if a web page refers to a geoscience dataset, the GeoLink web component will search GeoLink for any references to that dataset's URI and display the results on the page (Figure 5). Open Core Data,[17] an NSF-funded initiative focused on making data from continental and ocean drilling projects publicly available, currently uses GeoLink web components.

The GeoLink ontology itself is also of use to geoscientists. For example, the EarthCube Science Support Office (NSF Award 1623751) is building a reference implementation of the EarthCube architecture. This project leverages the philosophy of self-publishing semantic metadata to describe data repositories, capabilities, and services, as well as dataset-level metadata for discovery using schema.org markup to move towards the primary EarthCube goal of democratizing and improving access to data. GeoLink classes are employed to handle some of the ambiguity of the schema.org classes and properties.[18] Additionally, the Lamont-Doherty Earth Observatory has published two datasets related to physical samples, the Ocean Biogeographic Information System (OBIS)[19] and the Earthchem Database,[20] according to the GBO.

The above resources were all developed by members of the GeoLink project team, but external groups are using the data within GeoLink as well. For example, individuals

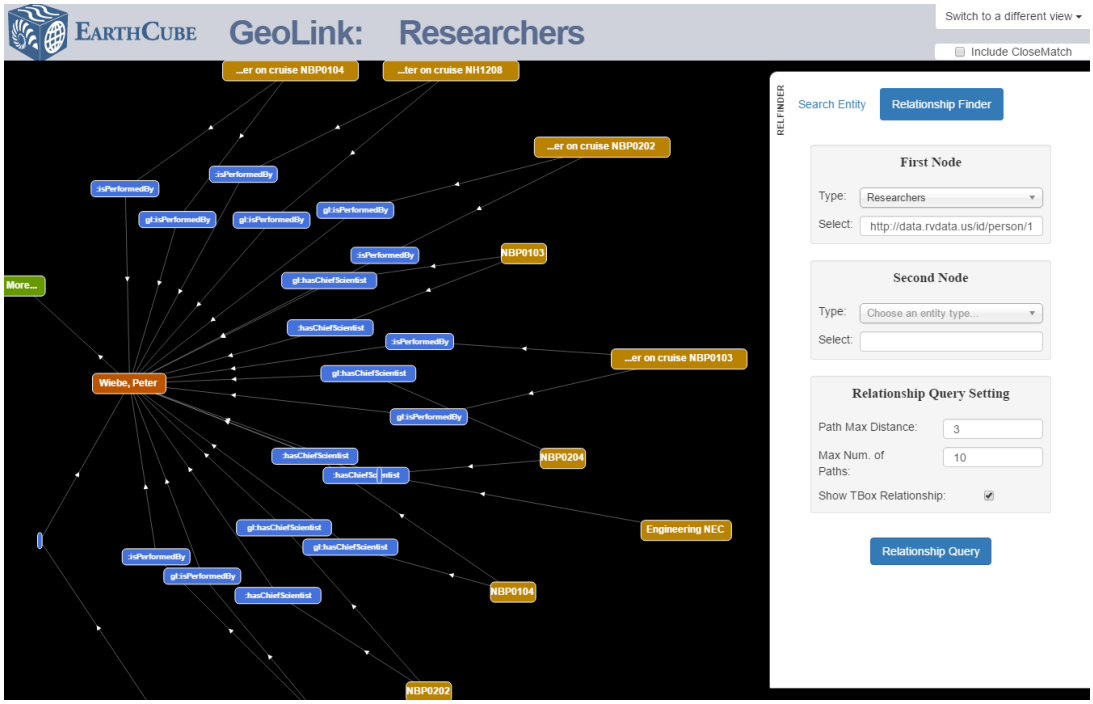

Figure 4. A web-based tool that allows users to explore the GeoLink linked data set.

from the NOAA National Centers for Environmental Information (NCEI), the Index to Marine and Lacustrine Geological Samples (IMLGS) and the CLIVAR & Carbon Hydrographic Data Office (CCHDO) have some information, such as that related to oceanographic cruises, that overlaps with data in GeoLink. These groups leverage that overlap to cross-check their own data for inconsistencies. As the GeoLink knowledge graph becomes well known within the geoscience research community, we expect the usage of the data by others external to the project to continue to grow.

4.4. Computer Science Usage

The GeoLink knowledge graph is the culmination of a three-year data integration effort that required intense manual effort by ontologists, data providers, and domain experts. Many computer scientists are currently working to automate some of the tasks involved in this type of effort in order to facilitate faster data integration. These automated systems fall into two groups: ontology (or schema) alignment and coreference resolution (or instance matching). It is important that researchers developing both types of systems have appropriate benchmarks on which to evaluate the performance of their automated tools on data integration tasks. Currently, the most widely accepted benchmarks are those within the Ontology Alignment Evaluation Initiative (OAEI).[21] These benchmarks were created beginning in 2004 and have spurred great innovation. However, some elements of real-world data integration challenges are not yet well represented within the OAEI benchmarks. The GeoLink knowledge graph can be used to fill in some of these gaps.

At the schema level, all of the OAEI benchmarks consist entirely of 1-to-1 equivalence relations (e.g. a Person in one dataset is equivalent to a Human in another). In practice, many of the relationships that exist between data sets are more complicated,

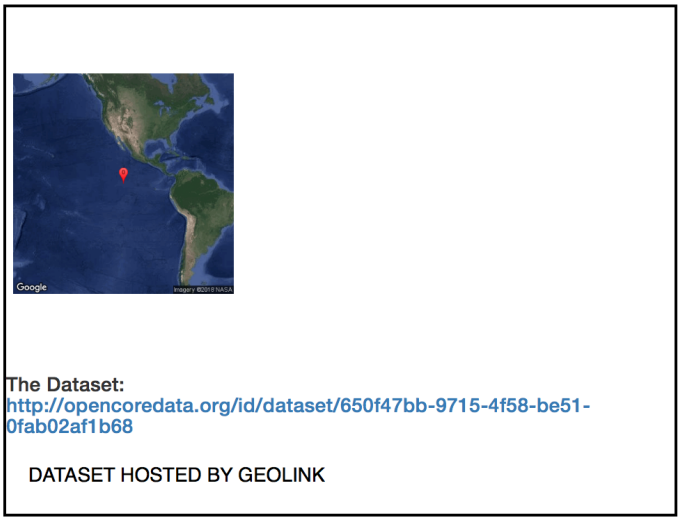

Figure 5. A GeoLink Web Component, which allows web developers to easily incorporate GeoLink data onto their sites.

such as a Professor with a hasRank property value of "Assistant" in one ontology is a subclass of the union of the Faculty and TenureTrack classes in another. Unfortunately, the majority of research activity in the field of ontology alignment remains focused finding only 1-to-1 equivalence relations. One reason may be that there are no widely accepted ontology alignment benchmarks that involve complex relations. As mentioned previously, GeoLink is published according to the GeoLink Base Ontology, which was derived from the GeoLink Modular Ontology. Because the GBO was manually engineered directly from the GMO and a SPARQL query was created to mitigate each change that was made, the alignment formed by these SPARQL queries is guaranteed to contain all of the relations necessary to solve this real-world alignment problem and no superfluous relations. We argue that this characteristic makes the GeoLink ontologies a good example of a complex ontology alignment problem that can be used as a benchmark. We have therefore made the alignment, along with the GBO and GMO, publicly available.[22]

Regarding coreference resolution, some of the benchmarks within the OAEI are synthetic, meaning that they were created by taking the individuals within a single dataset and modifying them. Coreference resolution systems are then judged on how accurately they can associate a modified individual with its original. This allows both precision and recall to be assessed. However, it is not clear that a strong performance on these synthetic benchmarks would carry over to real-world data integration tasks. The OAEI also contains some non-synthetic instance matching tasks, but these are focused somewhat narrowly on disambiguating authors based on information about their publications.

We have proposed the use of GeoLink as a way to expand the set of coreference resolution benchmarks, see Cheatham, Amini, and Patel (2016). Domain experts and data providers have manually established hundreds of coreferences among people and oceanographic cruises within GeoLink. These links are not all of the coreferences that exist, so recall cannot be assessed using this benchmark; however, precision on this real-world task can be effectively evaluated. Additionally, the GeoLink coreference resolution task has the potential to spur innovations related to scalability and align-

ing data with geospatial and temporal aspects, which are challenging areas for many existing systems.

5. Availability

The GeoLink knowledge graph is deployed at http://data.geolink.org and can be queried by both human and machine clients using SPARQL. In addition, the visualization interface described in Section 4.3 is available at http://demo.geolink.org. The GeoLink endpoint is maintained by NCEAS, which has been in existence since 1995 and is supported by organizations such as NASA, the NSF, and the National Academy of Sciences. In addition, a Creative Commons-licensed snapshot of the GeoLink knowledge graph is archived at the MBLWHOI Library, accessible via the citable URL http://hdl.handle.net/1912/9524. The DOI for this snapshot is 10.1575/1912/9524.

6. Conclusions and Future Work

This paper presented the GeoLink knowledge graph, a public and freely available source of geoscience data composed of seven of the largest data repositories in this domain. The structure of the knowledge graph and the data within it are described and links to more detailed information are provided to facilitate reuse. The quality and utility of the dataset are evidenced by the substantial amount of use that has already occurred in both the geosciences and computer science.

In the future, we hope to integrate additional geoscience data repositories into the knowledge graph. On the computer science side, we plan to put forth the alignment problem between the GMO and GBO as a potential new track within the Ontology Alignment Evaluation Initiative, in order to encourage the development of automated alignment systems capable of making large-scale integration projects like GeoLink easier in the future. Additionally, the GeoLink knowledge base has not yet been aligned to any upper level ontology, such as the Basic Formal Ontology (BFO). While such an ontology was not necessary for achieving the basic operational goals of the project, doing so could facilitate interoperability with other knowledge graphs. As a result, this is also an avenue of future work.

Acknowledgements

The authors sincerely thank their colleagues on the GeoLink team: Bob Arko, Suzanne Carbotte, Cyndy Chandler, Doug Fils, Yingjie Hu, Kerstin Lehnert, Bryce Mecum, Audrey Mickle, Lisa Raymond, Mark Schildhauer, and Peter Wiebe. The GeoLink project was supported by NSF Award No. 1440202.

Notes

[1]https://www.earthcube.org
[2]http://www.rvdata.us
[3]http://www.bco-dmo.org
[4]https://www.iodp.org
[5]http://www.mblwhoilibrary.org

[6]http://www.geosamples.org
[7]https://www.dataone.org
[8]https://minerals.usgs.gov/science/natl-geochemical-db/
[9]http://www.usap-dc.org
[10]http://www.nerc.ac.uk
[11]http://xmlns.com/foaf/spec/
[12]http://www.dublincore.org/specifications/
[13]https://www.w3.org/TR/prov-o/
[14]https://orcid.org
[15]https://www.doi.org
[16]http://geocomponents.org
[17]http://opencoredata.org
[18]https://github.com/earthcubearchitecture-project418/
[19]http://host.geolink.org/graph/obis
[20]http://host.geolink.org/graph/earthchem
[21]http://oaei.ontologymatching.org
[22]http://dase.cs.wright.edu/content/complex-alignment-benchmark-geolink-dataset

References

Berners-Lee, T. (2006). *Linked Data - design issues.* Retrieved from `http://www.w3.org/DesignIssues/LinkedData.html`

Cheatham, M., Amini, R., & Patel, C. (2016). Matching Instances in GeoLink. In *Proceedings of the 11th International Workshop on Ontology Matching* (Vol. 1766, pp. 237–238). ceur-ws.org.

Hitzler, P., Gangemi, A., Janowicz, K., Krisnadhi, A., & Presutti, V. (Eds.). (2016). *Ontology Engineering with Ontology Design Patterns: Foundations and Applications* (Vol. 25). IOS Press/AKA Verlag.

Krisnadhi, A., Hu, Y., Janowicz, K., Hitzler, P., Arko, R., Carbotte, S., ... others (2015a). The GeoLink Framework for Pattern-based Linked Data Integration. In *Proceedings of the posters and demos sesson at the 14th International Semantic Web Conference* (Vol. 1486). ceur-ws.org.

Krisnadhi, A., Hu, Y., Janowicz, K., Hitzler, P., Arko, R., Carbotte, S., ... others (2015b). The GeoLink Modular Oceanography Ontology. In *Proceedings of the 14th International Semantic Web Conference* (Vol. 9367, pp. 301–309). Springer.