A Complex Alignment Benchmark: GeoLink Dataset

Lu Zhou[1], Michelle Cheatham[1], Adila Krisnadhi[2], and Pascal Hitzler[1]

[1] DaSe Lab, Wright State University, Dayton OH 45435, USA,
{zhou.34, michelle.cheatham, pascal.hitzler}@wright.edu
[2] Faculty of Computer Science, Universitas Indonesia,
Depok, Jawa Barat 16424, Indonesia
adila@cs.ui.ac.id

Abstract. Ontology alignment has been studied for over a decade, and over that time many alignment systems and methods have been developed by researchers in order to find simple 1-to-1 equivalence matches between two ontologies. However, very few alignment systems focus on finding complex correspondences. One reason for this limitation may be that there are no widely accepted alignment benchmarks that contain such complex relationships. In this paper, we propose a real-world dataset from the GeoLink project as a potential complex alignment benchmark. The dataset consists of two ontologies, the GeoLink Base Ontology (GBO) and the GeoLink Modular Ontology (GMO), as well as a manually created reference alignment, that were developed in consultation with domain experts from different institutions. The alignment includes 1:1, 1:n, and m:n equivalence and subsumption correspondences, and is available in both EDOAL and rules syntax.

1 Introduction

Ontology alignment is an important step in enabling computers to query and reason across the many linked datasets on the semantic web. This is a difficult challenge because the ontologies underlying different linked datasets can vary in terms of subject area coverage, level of abstraction, ontology modeling philosophy, and even language. Due to the importance and difficulty of the ontology alignment problem, it has been an active area of research for over a decade [12].

Ideally, alignment systems should be able to uncover any entity relationships across two ontologies that can exist within a single ontology. Such relationships have a wide range of complexity, from basic 1-to-1 equivalence, such as a Person in one ontology being equivalent to a Human in another ontology, to arbitrary m-to-n relationships, such as a Professor with a hasRank property value of "Assistant" in one ontology being a subclass of the union of the Faculty and TenureTrack classes in another. Unfortunately, though, the majority of the research activities in the field of ontology alignment remains focused on the simplest end of this scale – finding 1-to-1 equivalence relations between ontologies.

Part of the reason for this may be that there are no widely used and accepted ontology alignment benchmarks that involve complex relations.

This paper seeks to take a step in that direction by proposing a complex alignment benchmark based on two ontologies which were developed by domain experts jointly with the reference alignment, and which in fact were developed for deployment on major ocean science data repository platforms, i.e., without the actual intention to develop an alignment benchmark. For this reason, the benchmark, including the reference alignment, can be considered to be (a) objective, in that it was created for deployment and not for benchmarking, (b) realistic, in that it captures an application use case developed for deployment, and (c) a valid ground truth alignment, in that the two ontologies and the reference alignment were developed together, by domain experts. We argue that it is therefore of rather unique nature and will inform complex ontology alignment research from a practical and applied, rather than artificial laboratory-like, perspective. The benchmark, coincidently, as this was the requirement of the use case, has a particular focus on relationships involving properties, which is particularly interesting because those have been shown to be rather difficult to handle for current alignment approaches [1].

The main contributions of this paper are therefore the following:

- Presentation of two ontologies to support data representation, sharing, integration, and discovery for the geoscience research domain.
- Creation of an alignment of these two ontologies that includes 1:1, 1:n, and m:n correspondences, and given the creation history and usage of the alignment, it is fair to say that the alignment constitutes a gold-standard reference.
- Publication of the benchmark alignment in both rule and EDOAL[3] syntax at a persistent URL[4] under a CC-BY license.

In addition, we have analyzed and categorized the mapping rules constituting the alignment. We found several which had not been classified or discussed previously, and we will present and discuss our analysis.

This paper is organized as follows. Section 2 discusses the few existing ontology alignment benchmarks that involve relationships other than 1-to-1 equivalence. Section 3 gives further background on the GeoLink modeling process, including why two different but related ontologies were developed. Section 4 discusses the alignment between the two GeoLink ontologies, along with some descriptive statistics and an analysis of the types of mapping rules constituting the alignment. Section 5 concludes with a discussion of potential future work in this area.

[3] http://alignapi.gforge.inria.fr/edoal.html
[4] http://doi.org/10.6084/m9.figshare.5907172

2 Related Work

Most work associated with evaluating the performance of ontology alignment systems has been done in conjunction with the Ontology Alignment Evaluation Initiative (OAEI)[5]. These yearly events allow developers to test their alignment systems on various tracks that evaluate performance on different facets of the problem such as instance matching, large ontology matching, and interactive matching, among others. Currently, most of these tracks involve the identification of 1-to-1 equivalence relationships, such as a Participant being equivalent to an Attendee. In 2009, the OAEI ran an "oriented" matching track that challenged systems to find subsumption relationships such as a Book is a subclass of a Publication. However, this track was abandoned after one year. Some system developers complained that the quality of the reference alignment was low [2]. This frustrated system developers and hindered participation. A discussion at the last two Ontology Matching workshops[6] made it clear that the community is interested in complex alignment, but that lack of applicable benchmarks is hindering progress. Our proposed benchmark seeks to address this concern by providing a reference alignment as a benchmark, and by addressing the quality issue of the previous benchmark by the fact that the process leading to the reference alignment guarantees its high quality.

In addition to using the OAEI benchmark, alignment systems that attempt to identify subsumption relations have sometimes used their own manually developed (and sometimes unpublished) reference alignments [5]. Other subsumption systems have evaluated the precision of their approach by manually validating relations produced by their system, while foregoing an assessment of recall [13]. Other related work has centered on developing a benchmark for compound alignments, which the authors define as mappings between class or property expressions involving more than two ontologies [10]. Their first step in this direction was to create a set of reference alignments containing relations of the form $< X, Y, Z, R, M >$, where X, Y and Z are classes from three different ontologies and R is a relation between Y and Z that results in a class expression that is related to X by the relation M. For example, a DisabledVeteran (X) is equivalent to (M) the intersection (R) of Veteran (Y) and Disabled (Z). This benchmark is based on cross-products among the OBO Foundry biomedical ontologies, which have been manually validated by at least two experts.

The work presented herein differs from these approaches by considering a wider range of relationship types (beyond subsumption and the type of ternary relation described in [10]), as they naturally arose out of the application from which the reference alignment was taken.

More related work is currently being undertaken by Thieblin and her colleagues, who are creating a complex alignment benchmark using the Conference track ontologies within the OAEI [14]. This work is partially completed, and at the time of this writing it covers three of the seven ontologies. In addition, we are

[5] `http://oaei.ontologymatching.org`
[6] `http://www.ontologymatching.org/`

collaborating with them (under their direction) to complete the dataset and prepare a new task in OAEI to evaluate complex alignment systems. The reference alignment we describe herein differs from the effort by Thieblin et al. in that the GeoLink ontologies and alignment constitute real-world datasets designed for practice and applied by geoscientists, rather than being an artificial artifact designed solely for alignment benchmarking. Furthermore, data from seven geoscience repositories have been published according to the GeoLink schema and they are available online[7]. This instance data can in the future be used by alignment systems that employ extensional matching techniques. In contrast to this, significant instance data is not readily available for most of the OAEI Conference Track ontologies.

3 The GeoLink Modeling Process

Benchmarks come in at least two varieties. On the one hand there are artificial benchmarks which provide a kind-of laboratory setting for evaluation. On the other hand there are benchmarks created from data as it is used in realistic use cases or even deployed scenarios. Both of these types are important, and they cover different aspects of the spectrum, and may have different advantages. Artificial benchmarks can be made to be balanced, or to focus on certain aspects of a problem, and sometimes they can be used to test scalability issues more easily as different versions of the same benchmark set may be easily producible. Natural benchmarks, on the other hand, may expose issues arising in practice which may easily be overlooked by designers of artificial benchmarks, in particular in a young field such as complex ontology alignment. Natural benchmarks also may come with an independently verified gold standard baseline, as in our case.

The project this benchmark arose from is called GeoLink [15] and was funded under the U.S. National Science Foundation's EarthCube initiative. This planned decade-long endeavor is a recognition that oftentimes the most innovative and useful discoveries come at the intersection of traditional fields of research. This is particularly true in the geosciences, which often bring together disparate groups of researchers such as geologists, meteorologists, climatologists, ecologists, archaeologists, and so on. For its part, GeoLink employs semantic web technologies to support data representation, sharing, integration, and discovery [9]. In particular, seven diverse geoscience datasets have been brought together into a single data repository.

At the beginning of the project, some providers' data resided in relational databases while others' had been published as RDF triples and exposed via a SPARQL endpoint. Because each provider had their own schema, the first step in the GeoLink project was to develop a unified schema according to which all data providers could publish their data [9]. Creating a unified schema for independently developed datasets is sometimes difficult, and the final product often ends up requiring providers to shoehorn their data into a schema that does not quite

[7] `http://data.geolink.org`

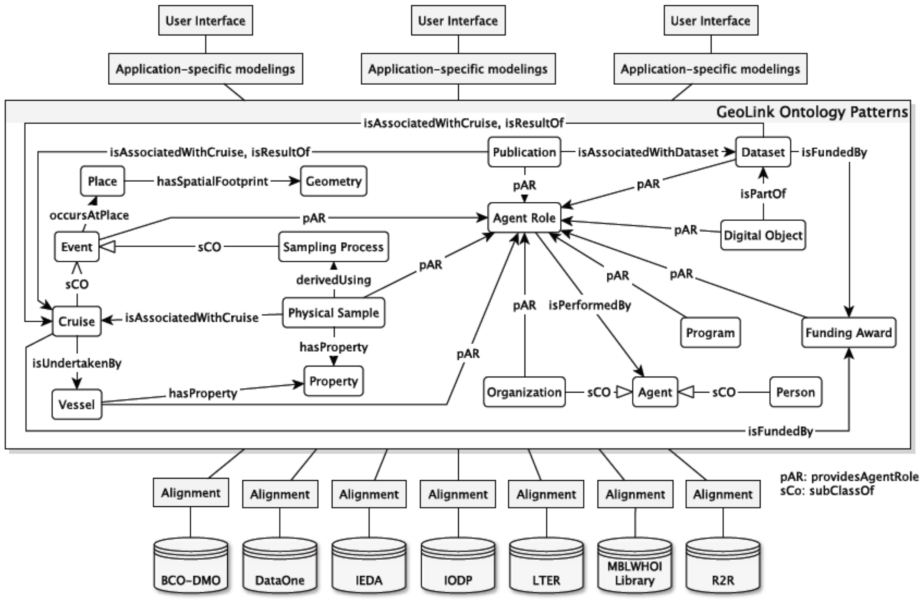

Fig. 1: Intended usage of the GMO

fit. GeoLink uses an approach that relies on ontology design patterns (ODP) in an attempt to avoid this issue [4]. An ODP represents a reusable solution to a recurring modeling problem. An ODP generally encodes a specific abstract notion, such as a process, event, agent, etc. These are frequently the small areas of semantic overlap that exist between datasets from different subfields of the same high-level domain. ODPs provide a structured and application-neutral representation of the key concepts within a domain. Throughout the first year of the project, geoscientists, data providers and ontologists worked together to identify and model the important concepts within the geosciences that recurred across two or more datasets. The result of this were what we call ontology modules, based on ODPs, and eventually they were stitched together to form the GeoLink Modular Ontology (GMO) [7].

As shown in Figure 1, the GMO allows data providers to publish only those aspects of their data modeled by the GMO according to that schema. Any data the provider has that is not covered by that schema can be published using the provider's own schema, since no other providers have similar content. For example, in the figure, the provider R2R has data related mostly to the cruise and vessel modules in the lower left of the figure, and so it publishes its related data using that terminology. R2R also has data not modeled by the GMO and so it uses its own terminology when publishing that information. This freedom is intended to make the publishing process easier; however, some problems still remained.

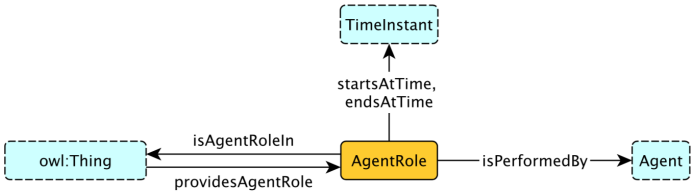

Fig. 2: The Agent Role pattern

Some of the patterns contain a rather complicated structure, mostly due to reification, which was employed to accommodate different perspectives (e.g., based on granularity) on the data. For example, many of the data providers have information about the sponsor of a project, and R2R has a native relation in their schema called hasSponsor with domain Award and range Organization. However following best practices, it leads to a more versatile model if being a sponsor is recognized (and thus modeled) as a role which an agent (in this case an organization) can assume. Creating a distinct relation for each type of role on a project (sponsor, chief scientist, research assistant, etc.) is brittle, in the sense that if new roles will be added later, potentially due to the inclusion of a new dataset, then the schema will need to be edited by adding new vocabulary for new roles together with (possibly complex) role relationships. Another issue with using a relation such as hasSponsor is that a more fine-grained data repository may have additional temporal information related to the sponsor role, and then it is not clear how to add this temporal information to the hasSponsor model without punning. Essentially, hasSponsor should better be expressed as a ternary relation between award, organization, and the type of relation (in this case, being a sponsor) expressed using an individual which can be reused in all sponsor relationships. In terms of ODPs, this is realized by reusing the Agent Role pattern, shown in abstract form in Figure 2. This approach both allows new roles to be added easily (by subclassing AgentRole) and supports temporal queries if desired.

Unfortunately, while the data providers recognized the utility of this modeling approach, they found it cumbersome to map their data to it. Looking at their own schemas, 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. A simpler interface for the data providers was therefore requested.

To accommodate this, a second ontology, together with a manual alignment between this ontology and the GMO, was created to bridge the gap via an intermediate schema that is “flatter” than the patterns and closer to the data providers’ own schemas, but still easy to align to the GMO modules because it has been developed directly out of the GMO. This ontology is referred to as the GeoLink Base Ontology (GBO). The providers publish their data according to

the GBO and then SPARQL construct queries which encode the alignment can be used to map data to the GMO. From the very beginning, it was intended that the data integration process would be based on manual, and thus high-quality, mappings between different schemas. As a consequence, ontology alignment systems were not employed to make these mappings, not even to inform human decisions. All mappings were established as a collaborative effort between the data repository providers, the domain experts, and the ontology engineers involved in the modeling and deployment process. Because the GBO was manually engineered directly from the GMO in order to serve this particular purpose, the alignment is guaranteed to be precisely the one intended by the developers. I.e. the alignment is guaranteed to contain all of the relations necessary to solve this real-world alignment problem and no superfluous relations have been included. 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 for systems that attempt to automate such alignment processes: While it is not a synthetic benchmark, it reflects complex alignment issues encountered in practice.

The example below illustrates the use of the GBO and its alignment to the GMO. In the GBO, there is a relation called hasSponsor with a domain that includes Award and range Organization. This mirrors many of the providers' existing schemas. Providers publish triples either according to the GMO schema (e.g., if they have temporal information), or according to the GBO schema.

```
x:award1 a               view:Award ;
         view:hasSponsor x:org1 .
x:org1   a               view:Organization .
```

Then, the GBO-oriented triples are converted into the GMO schema using this SPARQL construct:

```
PREFIX view: <http://schema.geolink.org/dev/view#>
CONSTRUCT {
  ?x    a                  :FundingAward ;
        :providesAgentRole _:bn1 .
  _:bn1 :isPerformedBy     ?y ;
        a                  :SponsorRole .
  ?y    a                  :Organization .
} WHERE {
  ?x    a                  view:Award ;
        view:hasSponsor    ?y .
  ?y    a                  view:Organization
}
```

Let us look at this by means of a schema diagram. In Figure 3, the three nodes and the two solid arrows indicate the graph pattern used to express the sponsoring organization role in the GMO. The dashed arrow is that is sometimes called a *shortcut* [8]. This shortcut (which is not part of the GMO) "flattens"

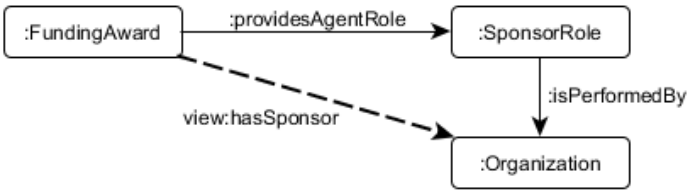

Fig. 3: A schema diagram to explain an example alignment

this part of the GMO, and in the GBO, the `:SponsorRole` node is removed, but the shortcut is added (and `:FundingAward` and `:Organization` have been replaced by the local `view:Award` and `view:Organization`, respectively).

Note that there is no doubt here about the intended alignment between the corresponding parts of the GBO and the GMO: `view:Award` and `:FundingAward` should be mapped to each other (as equivalent), as should `view:Organization` and `:Organization`. It is also clear that that the relation `view:hasSponsor` between an `view:Award` and an `view:Organization` should be aligned (as equivalence) to the concatenation of `:providesAgentRole` and `:isPerformedBy`, provided the entity shared by the two relation expressions is of type `:SponsorRole`, and the chain starts at a `:FundingAward` and ends at a `:Organization`. I.e. a complex alignment is required to express this very natural relationship between these two ontology snippets. Below we will give more examples of complex alignments arising from our setting, when we discuss the different alignment patterns we have identified. The example above is a "Typed Property Chain Equivalence" in our classification, and below we discuss this example further.

More information about the GMO and the project is available from [6] and from the project website[8].

4 The GeoLink Complex Alignment Benchmark

4.1 Dataset

In order to prepare the GeoLink ontologies for use as a complex alignment benchmark, some changes to the namespaces were required. As we introduced in the previous section, several ODPs and modules were created to represent the frequently recurring concepts in the GeoLink datasets, and these were stitched together to form the GeoLink Modular Ontology (GMO). During this process, the namespace of some entities was changed from one that reflected its originating pattern to the namespace of the GMO, which is `http://gmo#`. For example, the class FundingAward was originally in the fundingaward pattern, with the namespace `http://schema.geolink.org/1.0/pattern/fundingaward#`. After merging these modules, the namespace of the class FundingAward became `http:`

[8] `http://www.geolink.org/`

Ontology	Classes	Object Properties	Data Properties
GeoLink Base Ontology	40	149	49
GeoLink Modular Ontology	156	124	46

Table 1: The number of classes, object properties, and data properties in both GeoLink ontologies

`//gmo#`. This has been applied to all entities except those that are imported from other ontologies, which retain their original namespace. For example, the namespace of the class Instant, which is imported from `http://www.w3.org/2006/time#`, remains unchanged. Additionally, the namespace of entities in the GeoLink Base Ontology (GBO) has been changed from `http://schema.geolink.org/1.0/base/main#` to `http://gbo#`.

Table 1 shows the number of classes and properties in both ontologies. Both ontologies are comparable in size to ontologies currently used by the OAEI, meaning that they are within the capabilities of most current ontology alignment systems to handle.

4.2 Simple and Complex Correspondences

In order to understand the correspondences in the benchmark, we give the formal definition of simple and complex correspondences.

Simple Correspondence. Simple correspondence refers to basic 1-to-1 simple alignment between two ontologies, including class and property. It not only includes 1-to-1 equivalence, but also contains 1-to-1 subsumption, and 1-to-1 disjointness.

Complex Correspondence. Complex correspondence refers to more complex patterns, such as 1-to-n equivalence, 1-to-n subsumption, m-to-n equivalence, m-to-n subsumption, and m-to-n arbitrary relationship.

We have identified 12 different kinds of simple and complex correspondence patterns in the GeoLink complex alignment benchmark. Table 2 presents these different patterns and the corresponding number and category in the whole dataset. As the table shows, the alignment consists predominantly of complex relationships. In the following, we explain these alignment types, from simple 1-to-1 correspondence to complex m-to-n correspondence, with a formal pattern and example each.

Class Equivalence. The first pattern is just simple 1-to-1 class equivalence. Classes C_1 and C_2 are from ontology O_1 and ontology O_2, respectively.

Formal Pattern: $C_1(x) \leftrightarrow C_2(x)$
Example: $\mathsf{Award}(x) \leftrightarrow \mathsf{FundingAward}(x)$

Pattern	Occurrences	Category
Class Equivalence	10	1:1
Class Subsumption	2	1:1
Property Equivalence	7	1:1
Property Equivalence Inverse	5	1:1
Class typecasting Equivalence	4	1:n
Class typecasting Subsumption	1	1:n
Property typecasting Subsumption	5	1:n
Property typecasting Subsumption Inverse	5	1:n
Typed property chain Equivalence	26	m:n
Typed property chain Equivalence Inverse	17	m:n
Typed property chain Subsumption	17	m:n
Typed property chain Subsumption Inverse	12	m:n

Table 2: The alignment pattern types found in the GeoLink complex alignment benchmark, along with the number of times each occurs and the type of relation.

Class Subsumption. This pattern is very similar to the first pattern. But, instead of class equivalence, this pattern describes simple 1-to-1 class subsumption.

Formal Pattern: $C_1(x) \rightarrow C_2(x)$
Example: $\mathsf{GeoFeature}(x) \rightarrow \mathsf{Place}(x)$

Property Equivalence. Property alignment is also an important part of ontology alignment research [8]. This pattern captures simple 1-to-1 property equivalence. Property p_1 and property p_2 are from ontology O_1 and ontology O_2, respectively. The property can be either a data property or an object property.

Formal Pattern: $p_1(x, y) \leftrightarrow p_2(x, y)$
Example: $\mathsf{hasAward}(x, y) \leftrightarrow \mathsf{fundedBy}(x, y)$

Property Equivalence Inverse. This pattern is similar to the previous one, just that the domain and range values of a property are switched when it aligns to a property in another ontology.

Formal Pattern: $p_1(x, y) \leftrightarrow p_2(y, x)$
Example: $\mathsf{isAwardOf}(x, y) \leftrightarrow \mathsf{fundedBy}(y, x)$

Class Typecasting Equivalence. This pattern is more specific than the previous ones. The idea of typecasting, and why it is important in ontology mod-

eling, is formally introduced and discussed in [8]. The pattern indicates that individuals of type C_1 in one ontology are cast into a subclass of C_2 in the other ontology. Note that punning is employed here – x is treated as an individual on the left hand side of the rule and as a class on the right hand side. For example, an instance of PlaceType in the GBO might be 'ocean'. This is cast into a subclass of Place in the GMO. The reverse is also true: if the GMO has a subclass of Place called Island, then 'island' is an instance of the class PlaceType in the GBO.

Formal Pattern: $C_1(x) \leftrightarrow \mathsf{rdfs{:}subclassOf}(x, C_2)$
Example: $\mathsf{PlaceType}(x) \leftrightarrow \mathsf{rdfs{:}subclassOf}(x, \mathsf{Place})$

Class Typecasting Subsumption. This pattern is almost identical to the one above, except that the rule only holds in one direction. In the example, a GeoFeatureType (which comes from the General Bathymetric Chart of the Oceans[9] vocabulary) is always a type of Place, but there are types of Places that are not GeoFeatureType.

Formal Pattern: $C_1(x) \rightarrow \mathsf{rdfs{:}subclassOf}(x, C_2)$
Example: $\mathsf{GeoFeatureType}(x) \rightarrow \mathsf{rdfs{:}subclassOf}(x, \mathsf{Place})$

Property Typecasting Subsumption. This pattern is similar in spirit to the Class Typecasting patterns mentioned above. However in this case, a property is cast into a class assignment statement. In a sense, this alignment drops information, as y does not occur on the right hand side.

Formal Pattern: $p_1(x, y) \rightarrow \mathsf{rdf{:}type}(x, C_2)$
Example: $\mathsf{hasPlaceType}(x, y) \rightarrow \mathsf{rdf{:}type}(x, \mathsf{Place})$

We note here that some rules that fall under this category are not exact translations of the underlying SPARQL queries, due to expressibility constraints in EDOAL (see section 4.3 below). For instance, instead the example above, which states that the hasPlaceType object property is subsumed by an rdf:type statement with the range value of Place, we would actually like to state the following, which reflects the SPARQL query:

Formal Pattern: $p_1(x, y) \leftrightarrow \mathsf{rdf{:}type}(x, y) \wedge \mathsf{rdfs{:}subclassOf}(y, C_2)$
Example: $\mathsf{hasPlaceType}(x, y) \leftrightarrow \mathsf{rdf{:}type}(x, y) \wedge \mathsf{rdfs{:}subclassOf}(y, \mathsf{Place})$

For instance, we would like a rule that implies that the GBO statement hasPlaceType(Honolulu,Island) is equivalent to stating that Honolulu is a type of Island *and* that Island is a subclass of Place in the GMO. In other words, one of the individuals occurring as a property filler on the GBO side is cast into a class on the GMO side. At the same time, the other property filler on the GBO

[9] `https://www.gebco.net`

side is asserted to be an instance of this class. However, this is not possible because the statement requires a variable (y), and that is not supported by the core EDOAL language. The EDOAL specification does mention a pattern language that might enable this type of statement, but it does not appear to be fully supported at this time.

Property Typecasting Subsumption Inverse. This pattern is the same as the one above, except that the property fillers are flipped.

Formal Pattern: $p_1(x, y) \rightarrow \mathsf{rdf{:}type}(y, C_2)$
Example: $\mathsf{isPlaceTypeOf}(x, y) \rightarrow \mathsf{rdf{:}type}(y, \mathsf{Place})$

Again, in some cases we would actually like to state the following, which cannot be fully expressed in EDOAL, to the best of our knowledge:

Formal Pattern: $p_1(x, y) \rightarrow \mathsf{rdf{:}type}(y, x) \wedge \mathsf{rdfs{:}subclassOf}(x, C_2)$
Example: $\mathsf{isGeoFeatureTypeOf}(x, y) \rightarrow \mathsf{rdf{:}type}(y, x) \wedge \mathsf{rdfs{:}subclassOf}(x, \mathsf{Place})$

Typed Property Chain Equivalence. A property chain is a classical complex pattern that was introduced in [11]. This pattern captures the situation related to the hasSponsor property discussed in detail in Section 3. The pattern applies when a property, together with a type restriction on one or both of its fillers, in one ontology have been used to "flatten" the structure of the other ontology by short-cutting a property chain in that ontology. The pattern also ensures that the types of the property fillers involved in the property chain are typed appropriately in the other ontology. The formal pattern and example are shown below. The classes D_i and property r are from ontology O_1, and classes C_i and properties p_i are from ontology O_2.

Formal Pattern:

$$D_1(x_1) \wedge r(x_1, x_{n+1}) \wedge D_2(x_{n+1}) \leftrightarrow C_1(x_1) \wedge p_1(x_1, x_2) \wedge C_2(x_2) \wedge \cdots \wedge p_n(x_n, x_{n+1}) \wedge C_{n+1}(x_{n+1})$$

Example[10]:

$$\mathsf{Award}(x) \wedge \mathsf{hasSponsor}(x, z) \leftrightarrow \mathsf{FundingAward}(x) \wedge \mathsf{providesAgentRole}(x, y) \wedge \mathsf{SponsorRole}(y) \wedge \mathsf{performedBy}(y, z)$$

Note that in this and all following patterns, any of the D_i or C_i may be omitted (in which case they are essentially $\top$). Also, for the left-to-right direction, we assume that $x_2, \ldots x_n$ are existentially quantified variables.

[10] In contrast to the example discussed in Figure 3, we leave out `:Organization` and `view:Organization`, because it is possible, in principle, that a non-organization agent (e.g., an individual) may sponsor.

Typed Property Chain Equivalence Inverse. This pattern is the same as the one above, except that the property fillers are flipped.

Formal Pattern:

$$D_1(x_1) \wedge r(x_1, x_{n+1}) \wedge D_2(x_{n+1}) \leftrightarrow C_1(x_{n+1}) \wedge p_1(x_{n+1}, x_n) \wedge C_2(x_n) \wedge \cdots \wedge p_n(x_2, x_1) \wedge C_{n+1}(x_1)$$

Example:

$$\mathsf{Award}(z) \wedge \mathsf{isSponsorOf}(x, z) \leftrightarrow \mathsf{FundingAward}(z) \wedge \mathsf{provideAgentRole}(z, y) \wedge \mathsf{SponsorRole}(y) \wedge \mathsf{performedBy}(y, x)$$

Typed Property Chain Subsumption. This is identical to the Typed Property Chain Equivalence pattern except that the relationship only holds in one direction.

Formal Pattern:

$$D_1(x_1) \wedge r(x_1, x_{n+1}) \wedge D_2(x_{n+1}) \rightarrow C_1(x_1) \wedge p_1(x_1, x_2) \wedge C_2(x_2) \wedge \cdots \wedge p_n(x_n, x_{n+1}) \wedge C_{n+1}(x_{n+1})$$

Example:

$$\mathsf{Cruise}(x) \wedge \mathsf{hasChiefScientist}(x, z) \rightarrow \mathsf{Cruise}(x) \wedge \mathsf{providesAgentRole}(x, y) \wedge \mathsf{AgentRole}(y) \wedge \mathsf{performedBy}(y, z)$$

Typed Property Chain Subsumption Inverse. This pattern is the same as the one above, except that the property fillers are flipped.

Formal Pattern:

$$D_1(x_1) \wedge r(x_1, x_{n+1}) \wedge D_2(x_{n+1}) \rightarrow C_1(x_{n+1}) \wedge p_1(x_{n+1}, x_n) \wedge C_2(x_n) \wedge \cdots \wedge p_n(x_2, x_1) \wedge C_{n+1}(x_1)$$

Example:

$$\mathsf{Cruise}(z) \wedge \mathsf{isChiefScientistOf}(x, z) \rightarrow \mathsf{Cruise}(z) \wedge \mathsf{providesAgentRole}(z, y) \wedge \mathsf{AgentRole}(y) \wedge \mathsf{performedBy}(y, x)$$

In [11], four alignment types were identified, some of which are subsumed by ours. We do not at all claim that our classification above is exhaustive, but we consider it a refinement of the ones listed in [11]. We conjecture that there are many more important ones of relevance to other use cases. Mapping out the space of complex alignment types is, in our understanding, helpful for further research into complex alignment algorithms.

4.3 Format in EDOAL and Rule syntax

As mentioned previously, SPARQL construct queries are used to convert data published by the data providers according to the GBO into the schema described

in the GMO, because the GMO employs modeling practices that enhance extensibility and facilitate reasoning. However, most ontology alignment benchmarks are not formatted in SPARQL but rather according to the format provided by the Alignment API [5]. The standard alignment format is not expressive enough to capture complex relations. However, the Alignment API also provides a format called Expressive and Declarative Ontology Alignment Language (EDOAL) that can be used to express these types of relations. This format can be read and manipulated programmatically using the Alignment API, and is therefore very convenient for ontology alignment researchers. In addition, EDOAL is already accepted by the ontology alignment community. It has been used by others when proposing new alignment benchmarks [10] and [14], and we continue that approach here. Because EDOAL can be difficult for humans to parse quickly, we have also expressed the alignments in using a naive rule syntax. The rule presentation is not intended for programmatic manipulation, but rather to make it easier for humans to read and understand the alignments. Both versions of the alignment, along with the GBO and GMO ontologies, can be downloaded from `http://doi.org/10.6084/m9.figshare.5907172` under a CC-BY License. We have merged the two ontologies according to this reference alignment and used HermiT [3] to verify that there are no inconsistencies. The GeoLink website[11] contains detailed documentation of the dataset and provides users with more insights about the resource, such as all entities, patterns, and relationships between them in both ontologies.

5 Conclusion

Complex alignment has been discussed for a long time, but relatively little work has been done to advance the state of the art of complex ontology alignment. The lack of an available complex alignment benchmark may be a primary reason for the slow speed of development. In addition, most current alignment benchmarks have been created by humans for the sole purpose of evaluating alignment systems, and they may not always represent real-world cases. In this paper, we have proposed a complex alignment benchmark based on the real-world GeoLink dataset. The two ontologies and the reference alignment were designed and created by ontologists and geoscience domain experts to support data representation, sharing, integration and discovery. We take advantage of these ontologies to create a complex alignment benchmark. In our dataset, the alignments not only cover 1:1 simple correspondences, but also contain 1:n and m:n complex relations. All correspondences required to convert between the two ontologies (a key goal of ontology alignment) are guaranteed to be present, because one ontology was consciously created from the other, with SPARQL queries to mitigate each change. In addition, the alignment has been evaluated by domain experts from different organizations to ensure high quality. Moreover, the ontologies and alignments in both rule and EDOAL syntax have been published in FigShare with an open access license for reusability.

[11] `http://schema.geolink.org/`

As future work in this area, we plan to put forth this alignment problem as a potential new track within the OAEI. After that, based on participants' feedback, we will modify the reference alignment if necessary to perfect the benchmark by making it more convenient to use. This may involve, for example, making the alignment available in additional formats. Furthermore, we also plan to create an automated alignment system to tackle the alignment problem set forth by this benchmark.

Acknowledgments We would like to thank the geosciences data providers for sharing the data, and the domain experts for helping understand the concepts to create the ontologies and evaluate the reference alignment. In addition, we would also like to show our gratitude to Jerome Euzenat for providing advice regarding the conversion of rules to EDOAL.

References

1. Cheatham, M., Hitzler, P.: The properties of property alignment. In: Proceedings of the 9th International Workshop on Ontology Matching collocated with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Trentino, Italy, October 20, 2014. pp. 13–24 (2014)
2. David, J.: AROMA results for OAEI 2009. In: Proceedings of the 4th International Workshop on Ontology Matching (OM-2009) collocated with the 8th International Semantic Web Conference (ISWC-2009) Chantilly, USA, October 25, 2009 (2009)
3. Glimm, B., Horrocks, I., Motik, B., Stoilos, G., Wang, Z.: Hermit: An OWL 2 reasoner. J. Autom. Reasoning **53**(3), 245–269 (2014)
4. Hitzler, P., Gangemi, A., Janowicz, K., Krisnadhi, A., Presutti, V. (eds.): Ontology Engineering with Ontology Design Patterns - Foundations and Applications, Studies on the Semantic Web, vol. 25. IOS Press (2016)
5. Jain, P., Hitzler, P., Sheth, A.P., Verma, K., Yeh, P.Z.: Ontology alignment for linked open data. In: The Semantic Web - ISWC 2010 - 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I. pp. 402–417 (2010)
6. Krisnadhi, A.: Ontology Pattern-Based Data Integration. Ph.D. thesis, Wright State University (2015)
7. Krisnadhi, A., Hu, Y., Janowicz, K., Hitzler, P., Arko, R.A., Carbotte, S., Chandler, C., Cheatham, M., Fils, D., Finin, T.W., Ji, P., Jones, M.B., Karima, N., Lehnert, K.A., Mickle, A., Narock, T.W., 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, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part II. pp. 301–309 (2015)
8. Krisnadhi, A.A., Hitzler, P., Janowicz, K.: On the capabilities and limitations of OWL regarding typecasting and ontology design pattern views. In: Ontology Engineering - 12th International Experiences and Directions Workshop on OWL, OWLED 2015, co-located with ISWC 2015, Bethlehem, PA, USA, October 9-10, 2015, Revised Selected Papers. pp. 105–116 (2015)
9. Krisnadhi, A.A., Hu, Y., Janowicz, K., Hitzler, P., Arko, R.A., Carbotte, S., Chandler, C., Cheatham, M., Fils, D., Finin, T., Ji, P., Jones, M.B., Karima, N., Lehnert, K.A., Mickle, A., Narock, T., O'Brien, M., Raymond, L., Shepherd, A., Schildhauer, M., Wiebe, P.: The geolink framework for pattern-based linked data integration. In: Proceedings of the ISWC 2015 Posters & Demonstrations Track co-located

with the 14th International Semantic Web Conference (ISWC-2015), Bethlehem, PA, USA, October 11, 2015. (2015)
10. Pesquita, C., Cheatham, M., Faria, D., Barros, J., Santos, E., Couto, F.M.: Building reference alignments for compound matching of multiple ontologies using OBO cross-products. In: Proceedings of the 9th International Workshop on Ontology Matching collocated with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Trentino, Italy, October 20, 2014. pp. 172–173 (2014)
11. Ritze, D., Meilicke, C., Sváb-Zamazal, O., Stuckenschmidt, H.: A pattern-based ontology matching approach for detecting complex correspondences. In: Proceedings of the 4th International Workshop on Ontology Matching (OM-2009) collocated with the 8th International Semantic Web Conference (ISWC-2009) Chantilly, USA, October 25, 2009 (2009)
12. Shvaiko, P., Euzenat, J.: Ontology matching: State of the art and future challenges. IEEE Trans. Knowl. Data Eng. **25**(1), 158–176 (2013)
13. Suchanek, F.M., Abiteboul, S., Senellart, P.: PARIS: probabilistic alignment of relations, instances, and schema. PVLDB **5**(3), 157–168 (2011)
14. Thiéblin, É., Haemmerlé, O., Hernandez, N., dos Santos, C.T.: Towards a complex alignment evaluation dataset. In: Proceedings of the 12th International Workshop on Ontology Matching co-located with the 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 21, 2017. pp. 217–218 (2017)
15. You, J.: Geoscientists aim to magnify specialized web searching (2015)