

Complex alignment of modular ontologies

by

Reihaneh Amini

B.S., Alzahra University, 2013

M.S., Wright State University, 2016

M.S., Wright State University, 2017

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2025

Abstract

Ontology alignment, a critical process in the Semantic Web for detecting relationships between different ontologies, has traditionally focused on identifying so-called "simple" 1-to-1 relationships through class labels and properties comparison. The more practically useful exploration of more complex alignments remains a hard problem to automate, and as such is largely under-explored, i.e. in application practice it is usually done manually by ontology and domain experts. In this research, we explore the current state-of-the-art techniques in ontology alignment, focusing on several key areas. This project explores key challenges in ontology alignment, focusing on improving accuracy through enriched information, understanding performance differences between synthetic and real-world data, and evaluating how question formats and data representation affect user interaction and alignment outcomes. Recently, the surge in Natural Language Processing (NLP) capabilities, driven by advancements in Large Language Models (LLMs), presents new opportunities for enhancing ontology engineering practices, including ontology alignment tasks. This research explores the application of advanced language models to address complex challenges in knowledge representation and alignment. By adopting a prompt-driven methodology and incorporating rich, context-specific content modules, this study presents a novel framework aimed at enhancing automation in alignment processes. The proposed approach contributes to advancing intelligent systems capable of improving accuracy, scalability, and efficiency in knowledge integration tasks, with broad applicability across diverse domains such as information management, business intelligence, and decision support systems. Keywords: Complex Ontology Alignment · Ontology · Large Language Model · Knowledge Graph · Modular Ontology Modeling

Complex alignment of modular ontologies

by

Reihaneh Amini

B.S., Alzahra University, 2013

M.S., Wright State University, 2016

M.S., Wright State University, 2017

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
Carl R. Ice College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2025

Approved by:

Major Professor
Pascal Hitzler

Copyright

© Reihaneh Amini 2025.

Abstract

Ontology alignment, a critical process in the Semantic Web for detecting relationships between different ontologies, has traditionally focused on identifying so-called "simple" 1-to-1 relationships through class labels and properties comparison. The more practically useful exploration of more complex alignments remains a hard problem to automate, and as such is largely under-explored, i.e. in application practice it is usually done manually by ontology and domain experts. In this research, we explore the current state-of-the-art techniques in ontology alignment, focusing on several key areas. This project explores key challenges in ontology alignment, focusing on improving accuracy through enriched information, understanding performance differences between synthetic and real-world data, and evaluating how question formats and data representation affect user interaction and alignment outcomes. Recently, the surge in Natural Language Processing (NLP) capabilities, driven by advancements in Large Language Models (LLMs), presents new opportunities for enhancing ontology engineering practices, including ontology alignment tasks. This research explores the application of advanced language models to address complex challenges in knowledge representation and alignment. By adopting a prompt-driven methodology and incorporating rich, context-specific content modules, this study presents a novel framework aimed at enhancing automation in alignment processes. The proposed approach contributes to advancing intelligent systems capable of improving accuracy, scalability, and efficiency in knowledge integration tasks, with broad applicability across diverse domains such as information management, business intelligence, and decision support systems. Keywords: Complex Ontology Alignment · Ontology · Large Language Model · Knowledge Graph · Modular Ontology Modeling

Table of Contents

List of Figures	viii
Acknowledgements	ix
Dedication	x
1 Introduction	1
1.1 Introduction	1
1.2 Challenges and Advancements in Ontology Alignment	2
1.3 Outline	3
2 Knowledge Graph and Ontology Alignment	5
2.1 Overview	5
2.2 Ontology Alignment	6
2.3 Simple Vs. Complex Ontology Alignment	11
2.4 Interactive Ontology Alignment	13
2.5 Contribution	14
3 Ontology Alignment Evaluation Initiative	19
3.1 The Role of the Ontology Alignment Evaluation Initiative (OAEI)	19
3.2 Benchmarks And Evaluation In OAEI	20
3.3 Contribution	23
4 Complex Ontology Alignment Using Large Language Models	28
4.1 Overview	28

4.2	Contribution	32
5	Conclusion	34
5.1	Summary	34
5.2	Conclusion	35
	Bibliography	36
A	Contributions	38

List of Figures

2.1	Representation of “Micheal Cheadle” in the R2R database	7
2.2	Representation of “Micheal Cheadle” in the IODP database	7
2.3	Representation of “Micheal Cheadle” in the NSF database	8
2.4	Verification: Is “FundingAward” in GMO, same as “Award” is GBO?	16
2.5	Identification: What type of alignment exist between “FundingAward” in GMO, and “Award” is GBO?	16
3.1	GBO Cruise Ontology Schema	26
3.2	GMO Cruise Ontology Schema, (sCO=subClassOf and pAR=providesAgentRole)	26
4.1	Prompting workflow	31

Acknowledgments

I would like to express my deepest gratitude to my advisor, Prof. Pascal Hitzler, for his unwavering support throughout my research journey. Despite the distance, he was always just a message away, ready for a thoughtful brainstorming session. His patience, guidance, and immense knowledge have been invaluable, and I could not have asked for a better advisor.

I am also sincerely thankful to Prof. Michelle for her early mentorship during my transition from a Master's student to an aspiring PhD researcher. Her deep expertise, insightful guidance, and generous sharing of knowledge played a pivotal role in shaping my academic path. Combined with her unwavering encouragement, patience, and support, her expertise in the field laid a strong foundation for my research journey.

My heartfelt appreciation extends to all my colleagues and advisors who have accompanied me along the way, offering insights, collaboration, and encouragement.

A very special thank you to my husband, whose strong support has been my anchor, and to my parents, who inspired me to pursue this path and instilled in me the values of curiosity, perseverance, and academic excellence.

Dedication

To my father, who set me on this path.

Chapter 1

Introduction

1.1 Introduction

The World Wide Web was envisioned as a space equally accessible to both humans and machines; however, retrieving semantically relevant information remains a significant challenge. Linked Data addresses this issue by assigning unique, dereferenceable URIs to entities, facilitating standardized representation, interlinking, and cross-domain interoperability. This has led to the emergence of Linked Open Data, comprising billions of interconnected facts that support seamless knowledge integration. Within the Semantic Web, ontologies—particularly in graph-based form—serve as structured frameworks of entities and relationships. They provide a shared conceptual vocabulary that enables consistent data interpretation and exchange across stakeholders in a given domain.

Given its capacity to facilitate seamless knowledge sharing across domains and platforms, the adoption of linked data has grown substantially. Today, billions of facts spanning numerous subjects have been published as Linked Open Data (LOD)¹. Many of these datasets are indexed through platforms such as www.linkeddata.org, contributing to a continuously expanding LOD cloud. This interconnected ecosystem actively encourages data providers to link their own resources to existing datasets, thereby promoting interoperability and fostering the emergence of a globally integrated web of data.

An ontology is a structured representation of knowledge that defines entities and their relationships within a specific domain, enabling consistent data interpretation and interoperability. Ontology alignment is the process of identifying correspondences between entities across different ontologies, which is essential for integrating heterogeneous data sources and ensuring semantic consistency across systems. In this work, we explore the key challenges and strengths within the field of ontology alignment.

1.2 Challenges and Advancements in Ontology Alignment

Ontology alignment, a foundational task in the Semantic Web, has long posed significant challenges due to the complexity and diversity of ontological representations. A key issue arises in complex ontology alignment, where an entity in one ontology does not correspond to a single entity in another, but rather to a combination or transformation of multiple concepts. These complex correspondences are critical for enabling accurate and expressive data integration, especially in domains characterized by differing levels of abstraction, modeling granularity, or logical expressiveness.

Over the past decade, numerous systems have been proposed to address the complex alignment problem, yet systematic evaluation remains hindered by the lack of realistic benchmarks that reflect the intricacies of real-world data. Additionally, interactive ontology alignment—where human expertise is incorporated into the alignment process—has emerged as an important area of study. This raises key questions regarding the optimal methods for presenting ontological structures to users, whether through graphical depictions or RDF-based formats, in order to facilitate effective alignment identification.

Furthermore, recent advances in large language models (LLMs) offer new opportunities for enhancing ontology alignment tasks. While traditional systems may suffice for simple alignments, the capabilities of LLMs suggest significant potential for improving the detection of complex correspondences.

This thesis investigates these dimensions through three primary contributions:

- An analysis of how different forms of data presentation affect human performance in identifying ontology alignments;
- The development of a novel benchmark, GeoLink, designed to evaluate both simple and complex ontology alignment using real-world data;
- The design and implementation of a neural-symbolic system that leverages large language models to address complex ontology alignment challenges.

1.3 Outline

This thesis is a cumulative dissertation that presents foundational research on developing a complex ontology alignment system—one of the central challenges in the Semantic Web domain. As outlined in the introduction, several aspects of this field offer opportunities for advancing data integration through improved ontology alignment practices.

This research is organized into three key topics that collectively contribute to the overarching goal. The remainder of this dissertation is structured as follows: Chapter 2 introduces the first research topic, focusing on interactive ontology alignment practices. It examines how different methods of presenting ontologies to human users impact their ability to identify alignment tasks. The primary contribution of this chapter includes:

- Towards Best Practices for Crowdsourcing Ontology Alignment Benchmarks²

Chapter 3 focuses on the evaluation of ontology alignment systems within the OAEI initiative, highlighting the limitations of existing synthetic benchmarks. This chapter presents an in-depth analysis of the GeoLink Knowledge Graph and its Cruise Ontology, leading to the development of a new, realistic benchmark for OAEI. The primary contributions of this chapter include:

- GeoLink Cruises: A Non-Synthetic Benchmark for Co-Reference Resolution on Knowledge Graphs³

- The GeoLink knowledge graph⁴

Chapter 4 explores how recent advancements in Natural Language Processing, particularly the emergence of large language models (LLMs), are reshaping the future of ontology alignment systems. The chapter focuses on their application in addressing complex alignment tasks, highlighting the potential of LLMs to enhance alignment accuracy and scalability. The primary contributions of this chapter include:

- Towards Complex Ontology Alignment using Large Language Models⁵

Chapter 5 presents concluding remarks through a brief summary that highlights the overall contributions. Besides, we provide an outlook on future work.

Chapter 2

Knowledge Graph and Ontology Alignment

2.1 Overview

Tim Berners-Lee initially conceptualized the World Wide Web as an ecosystem equally accessible to both humans and machines⁶. While the modern web provides access to vast volumes of data, retrieving relevant and semantically related information about a specific concept remains a complex and often inefficient task for both users and automated systems. Even in the most straightforward cases, effective data discovery typically requires both background knowledge and advanced Natural Language Processing (NLP) capabilities. Despite these tools, it remains challenging to determine whether the results of a given query actually contain the precise information being sought.

The paradigm of Linked Data was introduced to address these challenges by defining a set of principles for the standardized representation, interlinking, and retrieval of data on the web. These principles include assigning a unique identifier, typically a Uniform Resource Identifier (URI), to each entity within a dataset. These URIs leverage the HTTP protocol to ensure that entities are both accessible and dereferenceable. When a URI is dereferenced, it provides not only descriptive information about the resource but also semantic links to

related entities⁶.

In the semantic web world, an ontology—particularly in its graph-based form—is defined as a ”coherent set of representational terms, together with textual and formal definitions, that embody a set of representational design choices”. It comprises a structured set of entities and the relationships among them, enabling domain experts to model and represent knowledge within a specific field. Fundamentally, an ontology serves as a shared conceptual framework or common language that allows stakeholders within the same domain to exchange and interpret data consistently⁷.

Ontologies are generally composed of two primary components:

Terminological Knowledge (T-Box): The T-Box defines the formal vocabulary of the domain, including its classes, properties, and logical rules⁷. It provides the structural schema or taxonomy upon which data is organized. For instance, a generic class such as Person may be included in the T-Box, along with constraints or axioms such as ”only a Person can possess the property hasFullName.”

Assertional Knowledge (A-Box): The A-Box contains concrete instances or assertions about individual entities in the domain. These assertions are constructed in accordance with the terminology and constraints specified in the T-Box. For example, an assertion such as ”Michael Cheadle is a Person” represents the assignment of an individual to a predefined class. Illustrative examples of such assertions are provided in Figures 2.1, 2.2, and 2.3 which is referring to the representation of the same entity in different datasets.

2.2 Ontology Alignment

In the Semantic Web context, an ontology is a structured and formal representation of knowledge within a specific domain. It consists of a coherent set of entities (such as classes, individuals, and properties) and the relationships among them, often expressed in a graph-based format. Ontologies serve as shared conceptual frameworks that enable consistent interpretation and communication of information between humans and machines. They embody specific design decisions that reflect how knowledge is modeled within a domain,

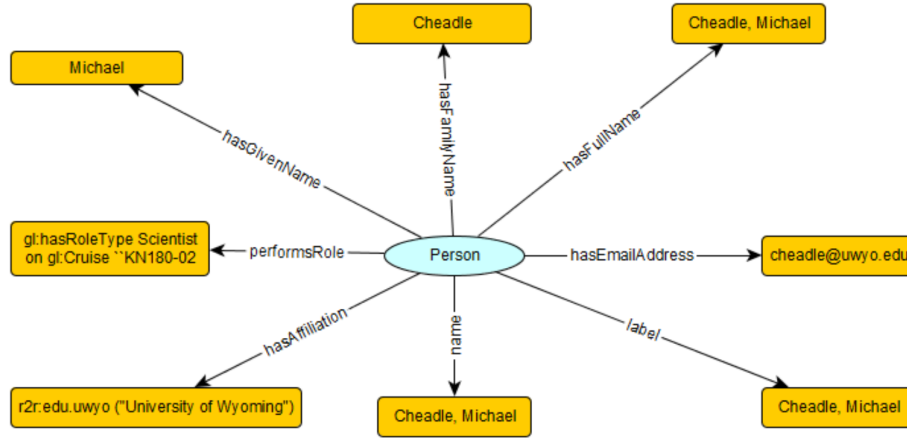


Figure 2.1: Representation of “Micheal Cheadle” in the R2R database

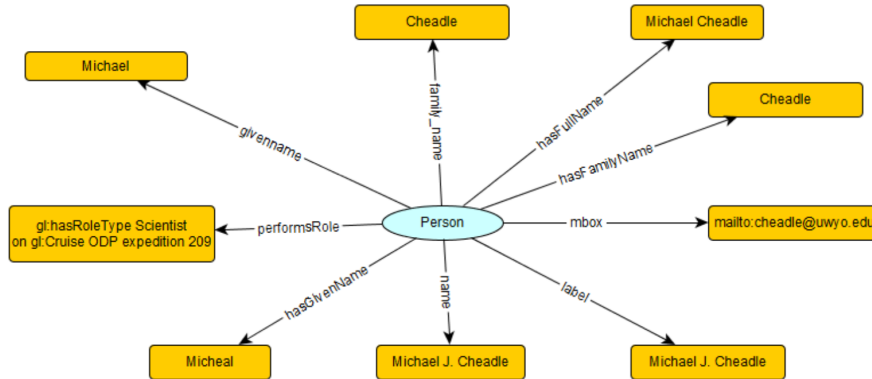


Figure 2.2: Representation of “Micheal Cheadle” in the IODP database

facilitating interoperability, data integration, and semantic reasoning.

Ontology alignment refers to the process of identifying correspondences between entities in different ontologies. These correspondences may be simple—such as equivalence between two classes—or complex, involving combinations or transformations of multiple entities. Alignment is essential when integrating heterogeneous datasets or enabling systems that rely on different ontological schemas to interoperate. Effective ontology alignment supports knowledge fusion, semantic search, and automated reasoning across diverse and distributed data sources.

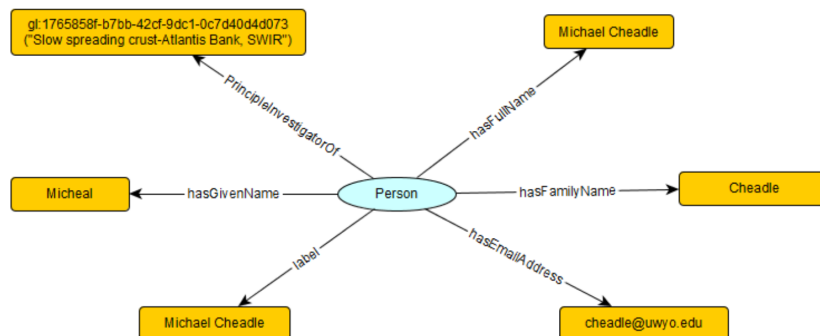


Figure 2.3: *Representation of “Micheal Cheadle” in the NSF database*

GeoLink¹ is a research initiative funded by the National Science Foundation (NSF) with the primary objective of integrating seven of the most significant oceanographic datasets in the United States. This integration is carried out in accordance with the principles of Linked Data, aiming to facilitate seamless interoperability across independently curated datasets. A central challenge in this endeavor is not merely the accessibility of distributed data, but the ability to integrate heterogeneous information into a semantically consistent and query-able whole.

To illustrate this, consider the example of a scientist named “Michael J. Cheadle,” whose information appears across three different data graphs: the R2R (Rolling Deck to Repository), IODP (International Ocean Discovery Program), and NSF datasets⁴. Each of these datasets contains partial and distinct information about the same individual. The R2R and IODP datasets record details about oceanographic research cruises where Michael Cheadle participated as a scientist, whereas the NSF dataset provides information about the research projects for which he served as principal investigator. Although the datasets refer to the same individual, the representation, attributes, and contextual associations vary between them.

This fragmentation exemplifies a common issue in distributed knowledge systems: accessing isolated data is not sufficient for extracting comprehensive insights. Effective data

¹<https://www.geolink.org/>

usage demands the semantic integration of diverse sources, raising two critical questions that must be addressed:

What is the meaning of the various data fields across different datasets?

How are these fields related to one another?

Without clear answers to these questions, data providers cannot efficiently share or consume data from other repositories. Even within the same domain, discrepancies in vocabulary, structure, and semantics hinder interoperability. Thus, data integration becomes a fundamental requirement, allowing users to interact with multiple datasets as if they constitute a single, unified repository.

Data integration within the context of ontologies operates at two primary levels: the Terminological Knowledge (T-Box) and the Assertional Knowledge (A-Box) levels.

At the A-Box level, integration focuses on identifying whether distinct data instances—often represented by different Uniform Resource Identifiers (URIs)—actually refer to the same real-world entity. This process is known as coreference resolution. For example, one dataset may include an individual named “Michael J. Cheadle,” while others record a “Michael Cheadle,” lacking the middle initial. A coreference resolution system would analyze available metadata (such as affiliations, email addresses, or unique identifiers) to assess whether these names denote the same person. Effective coreference resolution is critical for merging instance-level data and ensuring that queries over linked datasets return comprehensive and accurate results.

In contrast, data integration at the T-Box level pertains to aligning the schemas of different ontologies—a process referred to as ontology alignment. The objective of ontology alignment is to detect semantically equivalent or closely related concepts, classes, and properties across different ontologies. For example, the property `hasEmailAddress` in the R2R dataset may serve the same semantic purpose as the `mbox` property in the IODP dataset, both denoting an individual’s email address. Despite having different syntactic representations, these properties convey identical meaning, and recognizing this equivalence is crucial for enabling schema-level interoperability.

Ontology alignment plays a pivotal role in the broader goal of the Semantic Web and

Linked Data: to interconnect data from multiple sources in a meaningful and machine-readable way. It addresses the challenge of schema diversity—the variation in how different data providers model and describe their data—even within a shared domain of knowledge.

This thesis focuses on the problem of ontology alignment as a means of facilitating semantic integration across linked datasets. Ontology alignment systems typically take two ontologies as input and produce a set of correspondences between entities within these ontologies. Each correspondence includes:

- The URIs of the two entities being compared
- The nature of the relationship (e.g., equivalence, subsumption)
- A confidence score (ranging from 0 to 1) that quantifies the system’s certainty about the correctness of the match

Most alignment systems follow a structured pipeline. Since comparing every entity in one ontology to every entity in another is computationally expensive, a filtering step is often used to limit comparisons to potentially related entity pairs. For instance, aligning entities of type Person with entities of type Organization may be irrelevant and is thus typically excluded from the comparison process.

Following filtering, the system evaluates similarity metrics to assess candidate alignments. These metrics may include:

- **Syntactic similarity, which compares entity names or labels**
- **Semantic similarity, which draws upon lexical resources, ontologies, or knowledge graphs**
- **Structural similarity, which analyzes how entities are related to other entities within the ontology**

The outcome of this similarity assessment is a numerical score, which is then used to generate final matches. Typically, only those matches whose similarity scores exceed a predefined threshold are retained as valid alignments.

A comprehensive overview of ontology alignment techniques, including theoretical foundations and practical implementations⁸. The alignment of ontologies—particularly in domains with rich and diverse data such as oceanographic research—remains a critical enabler for advancing data interoperability, enhancing query capabilities, and promoting collaborative scientific discovery.

2.3 Simple Vs. Complex Ontology Alignment

Ontology alignment refers to the process of identifying semantic correspondences between entities—such as classes, properties, and instances—across two or more ontologies. These alignments play a fundamental role in enabling data interoperability across distributed and heterogeneous knowledge bases, particularly within the framework of the Semantic Web. By specifying how entities from different ontologies relate to each other, ontology alignment facilitates the integration, sharing, and reasoning over data drawn from diverse sources⁹.

Traditionally, most ontology alignment systems have focused on discovering simple alignments, which are typically one-to-one (1:1) relationships between individual entities. These simple correspondences usually represent equivalence (i.e., indicating that two entities are semantically identical) or, less frequently, hierarchical relationships such as subsumption (i.e., one concept being more general or specific than the other). The relative simplicity of these relationships makes them easier to identify using automated techniques based on lexical similarity, structural features, and statistical patterns. Furthermore, the evaluation of simple alignments is relatively straightforward, as it usually involves comparing predicted correspondences with a reference set and computing standard metrics such as precision and recall.

However, while simple alignments are effective in many basic use cases, they are limited in their ability to capture the full semantic complexity that often exists between different ontological representations. In real-world scenarios, especially those involving richly modeled domains, many semantic correspondences are inherently complex and cannot be adequately expressed through simple 1:1 mappings. These complex alignments may involve relationships

between combinations of entities, incorporate constraints, or reflect more nuanced semantic constructs. They are capable of representing richer semantic mappings, such as those involving conjunctions, disjunctions, cardinality conditions, and nested relationships among multiple ontology elements.

The need for complex alignments arises when the semantics of an entity in one ontology cannot be directly mapped to a single counterpart in another ontology, but instead corresponds to a composition or transformation of multiple concepts. These types of correspondences are essential for achieving a more precise and expressive integration of data, particularly in domains where ontologies are designed with varying modeling granularity, abstraction levels, or logical expressiveness.

Despite their importance, complex alignments have not received the same level of attention as their simpler counterparts in the development of alignment systems. The technical challenges associated with identifying, representing, and evaluating complex correspondences are significantly greater. Many current systems are either unable to detect such correspondences or are limited in their ability to handle them systematically. This gap results in alignment outputs that often underrepresent the true semantic relationships between ontologies, thereby limiting the effectiveness of data integration tasks.

Moreover, the standard evaluation methodologies for ontology alignment—typically based on exact matches with a gold-standard reference—tend to penalize systems that produce partially correct or near-miss complex alignments. These evaluation frameworks fail to account for degrees of correctness or the proximity of the generated alignment to the intended semantics. In practical settings, particularly within semi-automated alignment workflows, such near-miss correspondences may still hold considerable value, as they can be easily refined or confirmed by human experts. As such, evaluating alignment quality in binary terms (i.e., strictly correct or incorrect) often overlooks the potential contributions of complex alignment systems and provides limited diagnostic feedback for system improvement.

In conclusion, while simple alignments remain foundational in ontology alignment research and applications, they represent only a subset of the alignment challenges present in real-world scenarios. The development of robust methods for detecting, representing,

and evaluating complex alignments is critical for advancing semantic interoperability across ontologies. Future research must address both the algorithmic and evaluative aspects of complex alignment to ensure that alignment systems are equipped to support the full richness of semantic relationships encountered in diverse and evolving data ecosystems.

2.4 Interactive Ontology Alignment

In a wide range of domains, human cognitive abilities continue to surpass computational systems, particularly in tasks involving pattern recognition, contextual interpretation, and semantic alignment. Ontology alignment is one such task where humans often demonstrate superior performance. Humans can intuitively recognize conceptual equivalences, detect nuanced semantic relationships, and resolve ambiguities in meaning—capabilities that remain challenging for even the most sophisticated computational models. Consequently, many ontology alignment systems have been developed to incorporate human expertise at various stages of the alignment process¹⁰.

These systems range along a spectrum from fully manual to semi-automated approaches. Fully manual systems rely entirely on human judgment to identify correspondences between ontological elements, but such methods are impractical when dealing with large-scale ontologies, which may contain thousands of entities and properties. The time, effort, and cognitive load required for purely manual alignment make them infeasible for modern, data-intensive applications. As a result, contemporary ontology alignment efforts mostly focus on semi-automated systems that aim to balance the efficiency of machine algorithms with the accuracy and intuition of human users.

Semi-automated systems typically rely on computational techniques to generate candidate alignments between concepts in different ontologies. Human input is then incorporated at critical junctures—either to confirm the validity of proposed alignments or to assist in resolving cases where automated algorithms fail to reach a confident decision. One of the most prevalent approaches in this context involves generating a complete set of alignment suggestions through automated methods, followed by a human verification phase in which

users review and validate the proposed matches.

To enhance the scalability and efficiency of this verification step, optimization techniques have been introduced. One such strategy involves clustering similar alignment suggestions and presenting only a representative example from each cluster for human review, thereby reducing redundancy and minimizing user workload. Another common optimization is to prioritize human input for borderline cases—those matches whose computed similarity scores are near the decision threshold—on the assumption that such cases are more likely to be incorrect or uncertain.

Despite the clear advantages of human involvement in alignment tasks, a significant challenge lies in accessing individuals with the necessary domain-specific knowledge and expertise to perform these tasks accurately and efficiently. Domain experts and ontology engineers possess the ideal skill sets to support high-quality alignment, but in practice, they are often constrained by time limitations and competing professional obligations, making their availability for extensive data integration tasks quite limited.

To mitigate this issue, crowdsourcing has emerged as a promising strategy. By distributing alignment tasks to a large pool of contributors—often drawn from the general public or through online platforms—crowdsourcing enables the parallelization of effort and reduces the dependency on a small group of specialists. While the quality of contributions from non-experts may vary, careful task design, redundancy mechanisms, and aggregation techniques can be employed to ensure the reliability of crowdsourced results. Moreover, crowdsourcing approaches can be particularly effective for ontology alignment tasks that involve straightforward decisions or when aided by intuitive user interfaces and contextual guidance.

2.5 Contribution

During the early phase of my research in this field, I published a paper² that critically examines the reliability and design implications of utilizing crowdsourcing for developing ontology alignment benchmarks. Unlike prior studies that primarily aim to enhance crowdsourcing efficiency—such as minimizing time and cost—this work focuses on how task design and

formulation influence the quality and credibility of the resulting alignment data.

Specifically, the paper investigates four key research questions:

- Whether offering more expressive answer choices beyond binary (yes/no) responses enhance worker accuracy?
- The influence of question type (true/false vs. multiple choice) on the accuracy of crowd responses.
- The most effective methods for presenting contextual information to crowd workers to support accurate decision-making.
- The feasibility of identifying low-quality or fraudulent contributors (scammers) in ontology alignment microtasks.

By addressing these questions, we contributed to the broader discussion of how to ensure that crowdsourced ontology alignment results are robust, valid, and acceptable to the research community². The findings offer actionable insights for designing more effective crowdsourcing tasks and benchmarks in semantic web research. In summary, while the complexity of ontology alignment presents significant computational challenges, the integration of human judgment—either through expert verification or scalable crowdsourcing mechanisms—remains a vital component of effective alignment systems. Semi-automated methods that strategically incorporate human input are increasingly favored for their ability to balance scalability with semantic precision, ensuring the integrity and usability of integrated knowledge bases.

In the domain of ontology alignment, crowdsourcing has emerged as a promising approach to enhance the accuracy and scalability of alignment verification tasks. Previous efforts leveraging crowdsourcing methodologies have generally adopted two principal strategies for eliciting judgments about the semantic relationship between entities from users:

- (1) binary true/false questions
- (2) multiple-choice questions that capture more nuanced semantic relationships.

In the first approach, users are presented with a straightforward true/false prompt, typically phrased as a yes-or-no question such as: "Can 'Award' be matched with 'Funding Award'?" as shown in Figure 2.4. The user is then asked to simply indicate whether the two entities can be considered equivalent. This format is efficient and easy to process, making it well-suited for high-volume tasks or for engaging participants without deep domain expertise. However, it inherently restricts the expressive power of the alignment task, as it does not allow users to specify alternative forms of relationships beyond strict equivalence.

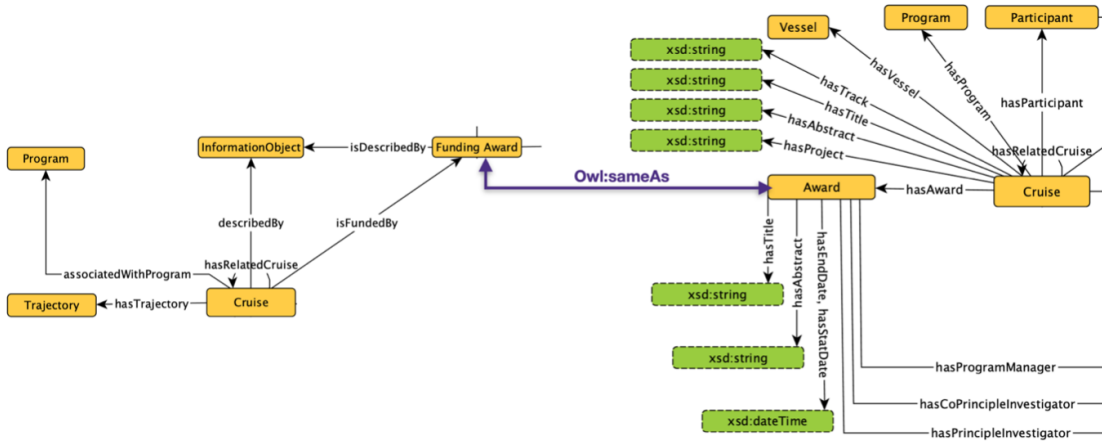


Figure 2.4: Verification: Is “FundingAward” in GMO, same as “Award” is GBO?

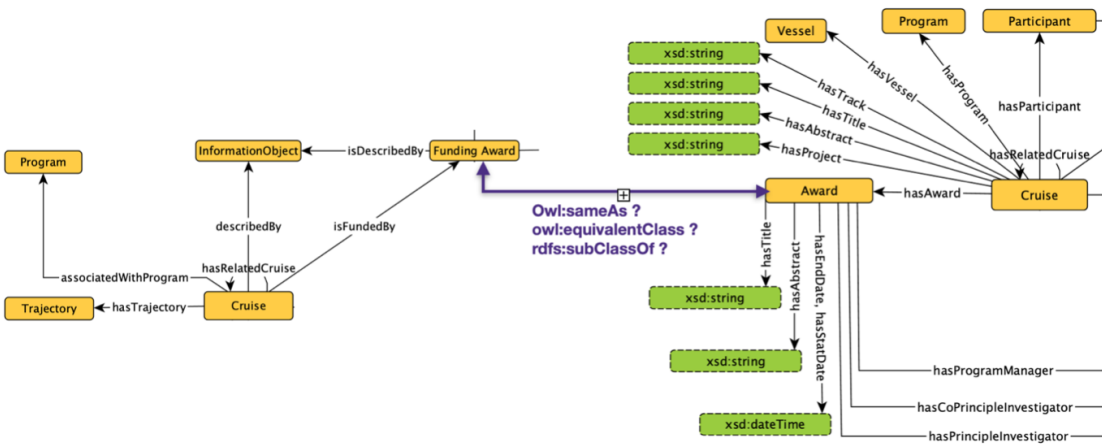


Figure 2.5: Identification: What type of alignment exist between “FundingAward” in GMO, and “Award” is GBO?

The second approach involves a multiple-choice format designed to elicit more detailed

semantic insights from users. For example, instead of asking whether “Award” and “Funding Award” are equivalent, the system might ask: “What is the relationship between ‘Award’ and ‘Funding Award’?”, ; as shown in Figure 2.5. Users would then choose from a set of predefined options, such as:

- “Paper and Contribution are the same” (indicating exact equivalence)
- “Every Paper is a Contribution, but not every Contribution is a Paper” (indicating a subclass relationship)
- “Every Contribution is a Paper, but not every Paper is a Contribution” (indicating a superclass relationship)
- “There is no relationship between Paper and Contribution.”

This more expressive question format supports a broader range of alignment types, including non-equivalence relationships such as subsumption, and enables a richer understanding of the semantic connections between ontology elements. Additionally, the use of clearly differentiated response options helps guide the user toward the most appropriate interpretation of the relationship, even in cases where the semantic distinctions may be subtle.

The motivation for adopting the multiple-choice format is closely aligned with the evolving needs of ontology alignment systems. As these systems advance beyond identifying simple one-to-one equivalence mappings and begin targeting more complex alignments—such as subclass relations, part-of hierarchies, and composite mappings—human input must also evolve to accommodate this increased complexity. Crowdsourcing platforms, therefore, must be designed to elicit granular semantic judgments that go beyond binary decisions.

Furthermore, modern semi-automated alignment pipelines often operate by first generating a pool of candidate matches for each entity within an ontology. These candidates are typically selected based on lightweight heuristics or lexical similarity measures. Subsequently, more computationally intensive methods—such as structural or semantic similarity computations—are applied to refine the candidate list and determine which correspondences are valid. Within this context, crowdsourcing can serve as an intermediate verification layer.

Specifically, the multiple-choice approach can be used to cast a wider net for potentially related entities and then leverage human insights to narrow down the most semantically appropriate matches. The filtered results, enriched by user feedback, can then be passed to alignment algorithms for further processing, improving both the precision and contextual accuracy of the final mappings.

In summary, while true/false questions offer simplicity and speed, multiple-choice formats provide the semantic depth necessary for addressing the growing complexity of ontology alignment tasks. Designing crowdsourcing systems that balance ease of use with the ability to capture nuanced relationships is essential for enabling scalable and effective human-in-the-loop alignment systems.

Chapter 3

Ontology Alignment Evaluation Initiative

3.1 The Role of the Ontology Alignment Evaluation Initiative (OAEI)

As the number and diversity of ontology alignment systems continue to grow, there is an increasing need for a standardized and systematic framework to assess, compare, and improve these systems. The Ontology Alignment Evaluation Initiative (OAEI) ¹ was established to address this need. It is a coordinated international effort that has been conducting annual evaluations of ontology alignment systems since 2004.

The OAEI serves several critical functions in the ontology alignment research community. Its primary objectives include:

- Evaluating the performance of ontology alignment systems across diverse tasks and domains,
- Facilitating collaboration and knowledge exchange among system developers,

¹<https://oaei.ontologymatching.org/>

- Driving improvements in the design, scalability, and accuracy of alignment methodologies, and
- Providing a unified, repeatable evaluation framework through standardized datasets, test cases, and benchmarking protocols.

To accommodate the varying requirements and complexities of alignment tasks, the OAEI organizes its evaluations into multiple tracks, each focusing on a specific type of alignment problem. These include tracks such as Ontology Alignment for Query Answering, Instance Matching, Interactive Matching Systems, among others. Each track provides domain-specific ontologies and evaluation criteria, enabling the fair assessment of systems under different conditions and application scenarios.

Results from these evaluations demonstrate that while a number of ontology alignment systems achieve high levels of performance and accuracy, none have yet attained perfect precision and recall. In other words, even the most advanced systems either fail to identify some valid correspondences (false negatives) or mistakenly include incorrect ones (false positives). This highlights the ongoing challenge in achieving comprehensive and error-free alignment, and underscores the importance of continued research and collaboration within the community.

The OAEI thus plays a vital role not only in benchmarking current capabilities but also in guiding the future development of more robust, adaptive, and semantically aware alignment systems.

3.2 Benchmarks And Evaluation In OAEI

The Ontology Alignment Evaluation Initiative (OAEI) has long served as the primary platform for researchers in the fields of ontology alignment and coreference resolution to evaluate, compare, and showcase their systems. Among the various tracks organized by OAEI, one notable effort is the evaluation of coreference resolution systems, which was formally introduced in 2004. Since its inception, this track has attracted a number of participants

annually, contributing to a growing body of empirical evidence regarding the effectiveness and scalability of such systems.

Central to this track is the Semantic Publishing Instance Matching Benchmark (SPIMBENCH)—a synthetic benchmark specifically designed to evaluate the ability of coreference resolution systems to identify instances referring to the same Creative Work¹¹. SPIMBENCH facilitates controlled and reproducible testing by providing a rich and scalable framework for synthetic data generation. It supports systematic variation through a wide range of transformations that simulate real-world challenges in ontology alignment while also ensuring a complete and reliable gold standard for evaluation.

The schema used by SPIMBENCH is reasonably complex, comprising 22 classes, 85 object properties, and 31 data properties. Synthetic datasets are created by applying various transformations to a base source file. These transformations fall into three primary categories:

- Value Transformations, which involve surface-level syntactic changes such as typographical errors, use of abbreviations, or substitution with synonyms;
- Structural Transformations, which alter the shape and depth of the ontology representation through mechanisms like property nesting or aggregation;
- Semantic Transformations, which manipulate schema-level relationships using constructs from OWL (Web Ontology Language), such as class equivalence, disjointness, or more complex class expressions.

The purpose of these transformations is to create realistic alignment challenges by generating a controlled set of synthetic matches and non-matches. However, because SPIMBENCH data is synthetically generated, it does not include coreferences that naturally emerge from independently curated datasets. Instead, coreferences are introduced programmatically, and the benchmark is explicitly designed to simulate plausible alignment tasks rather than mirror the full unpredictability of real-world data.

SPIMBENCH is available in two configurations: a small-scale version referred to as the sandbox, and a medium-scale version known as the mainbox. One of its key advantages is the flexible scalability it offers. Because the data is synthetically constructed, SPIMBENCH can generate large volumes of test cases to evaluate system performance under increasing data loads, making it highly suitable for scalability testing.

The availability of complete reference alignments within SPIMBENCH allows for the precise calculation of evaluation metrics such as precision and recall. Moreover, the benchmark is designed in a way that allows researchers to isolate specific types of transformations—be they semantic, structural, or syntactic—and assess how coreference resolution systems perform in response to each. This fine-grained diagnostic capability is one of the most valuable aspects of SPIMBENCH.

Nonetheless, SPIMBENCH also presents certain limitations. Its synthetic nature, while beneficial for controlled evaluation and repeatability, inherently narrows its scope. The benchmark assumes specific alignment patterns—often relying on string similarity or clearly defined structural changes—which may not adequately reflect the variety and ambiguity encountered in real-world data. For example, two instances such as "paint" and "color" may be semantically equivalent without exhibiting any string similarity. Likewise, an entity labeled "Organizer" may represent a Person in one dataset and a Company in another, depending on context and domain modeling—scenarios that synthetic benchmarks struggle to emulate convincingly.

Another point of critique is that SPIMBENCH employs a single ontology across both source and target datasets. While this simplifies evaluation, it does not capture the heterogeneity and schema mismatch that commonly arise in real-world data integration tasks. Real datasets often involve ontologies designed independently, with differing conceptualizations, terminologies, and modeling styles. This lack of ontological diversity limits the benchmark's ability to assess system robustness in realistic alignment scenarios.

Although SPIMBENCH incorporates some semantic complexity through OWL axioms and RDF Schema (RDFS) constructs, it still lacks the breadth of OWL features found in rich ontologies. Important constructs such as `owl:topDataProperty`, `owl:bottomDataProperty`,

owl:Thing, owl:Nothing, and owl:real are either underrepresented or entirely absent, reducing the generalizability of benchmark results. Including these constructs would significantly improve the realism and diversity of the benchmark, thereby strengthening its capacity to evaluate system performance under more demanding semantic conditions.

Similar limitations can be found in other synthetic instance-matching benchmarks independent of OAEI. For example, ONTOBI, which generates ontologies using Wikipedia infoboxes, also produces datasets of moderate size and complexity. Its schema consists of 17 classes, 13 object properties, and 128 data properties. However, as with SPIMBENCH, both the source and target datasets in ONTOBI are drawn from the same domain and employ the same ontology, thereby reducing heterogeneity and failing to fully reflect real-world challenges.

In conclusion, synthetic datasets such as SPIMBENCH provide invaluable tools for benchmarking and diagnosing coreference resolution systems. Their controlled nature allows for reproducible testing, scalability evaluation, and fine-grained performance analysis. These features are particularly useful in the early stages of system development, where systematic validation is essential. However, synthetic datasets alone cannot capture the full complexity and variability of real-world data. Therefore, while synthetic benchmarks should continue to be used for initial research and controlled experiments, they must be complemented by real-world datasets to ensure that alignment systems are equipped to handle the rich, ambiguous, and often unpredictable nature of semantic data integration at scale.

3.3 Contribution

The GeoLink Cruise benchmark³, in contrast to SPIMBENCH, exemplifies such real-world complexity. Its schema and A-Box (assertional data) are significantly more intricate, and the source and target datasets are based on independently developed and domain-specific ontologies. This introduces real ontological diversity and structural inconsistency, making it a more representative environment for testing coreference resolution and ontology alignment systems under realistic conditions.

Having discussed the advantages and limitations of synthetic benchmarks, we now turn our attention to the value and strengths of real-world benchmarks, and how they can provide more comprehensive insights into the performance of ontology alignment and instance matching systems.

The benchmark proposed in this study is derived from the GeoLink knowledge graph⁴, a real-world dataset that exemplifies the complexity and heterogeneity typically encountered in large-scale semantic integration efforts. GeoLink was developed as part of the National Science Foundation’s EarthCube initiative, with the objective of integrating data from seven distinct and diverse geoscience repositories into a unified semantic framework.

GeoLink’s integrated ontology is publicly available and provides a shared schema that brings together the conceptual models from each contributing dataset. The corresponding knowledge base can be queried via SPARQL endpoints, facilitating direct access to the structured data. As of the current version, the GeoLink knowledge graph comprises 282 classes, 338 properties, over 5.1 million instances, and approximately 45 million RDF triples. This scale and richness of data provide a valuable testbed for evaluating instance matching and ontology alignment approaches in real-world conditions, where data diversity, incompleteness, and semantic ambiguity are prevalent.

The primary objective of the GeoLink project was to integrate geoscience data from multiple repositories into a unified knowledge graph. To support this, GeoLink provides two ontologies: the GeoLink Modular Ontology (GMO) and the GeoLink Base Ontology (GBO).

The GMO is constructed using ontology design patterns (ODPs), ensuring a modular and semantically rich schema. Core concepts—such as Person or Physical Sample—are encapsulated as independent modules, developed collaboratively by domain experts, ontologists, and data providers. These modules form the foundation of the GMO, enabling data providers to publish RDF data aligned with a shared vocabulary. However, due to the complexity of some patterns, such as the AgentRole class, mapping can be challenging.

To address this, the GBO was introduced as a simplified version of the GMO. While it does not adhere to the same rigorous ontology design standards, it offers a more accessible schema for data publication. Both ontologies are publicly available under a CC-BY license.

Importantly, the GMO and GBO have also been used to construct a complex ontology alignment benchmark, highlighting the nuanced relationships between the two ontologies. This benchmark has been included in the Ontology Alignment Evaluation Initiative (OAEI) since 2018.

Cruise is one of the three core class types in the GeoLink knowledge base, alongside People and Organization. To support the evaluation of instance matching systems in a realistic setting, we introduce the GeoLink Cruise Benchmark³, which consists of two real-world Cruise datasets sourced from R2R² and BCO-DMO³.

This benchmark offers several distinguishing features that set it apart from existing benchmarks. Notably, it is derived from datasets curated and published by seven geoscience data providers as part of the GeoLink project. These datasets are inherently natural and domain-specific, reflecting real-world complexities, unlike synthetic benchmarks that often rely on simplified assumptions. As discussed previously, synthetic datasets are valuable for controlled testing and early-stage research, but they may introduce biases due to artificial assumptions and fail to represent the nuanced challenges present in real-world data integration.

A key advantage of the GeoLink Cruise Benchmark is the availability of a high-quality ground truth. Domain experts manually identified and curated 491 owl:sameAs links between matching Cruise instances in the R2R and BCO-DMO datasets. This manually verified alignment allows for robust evaluation of instance matching systems using standard metrics such as precision and recall. Additionally, the benchmark is scalable; further Cruise data can be readily retrieved from the GeoLink SPARQL endpoint, supporting extended experiments on scalability and system robustness.

Overall, the GeoLink Cruise Benchmark provides a realistic, semantically rich testbed for evaluating the practical effectiveness of instance matching systems in a real-world geoscience context.

Figures 3.1 and 3.2 illustrate the full set of classes and properties associated with the

²<http://www.rvdata.us>

³<http://www.bco-dmo.org>

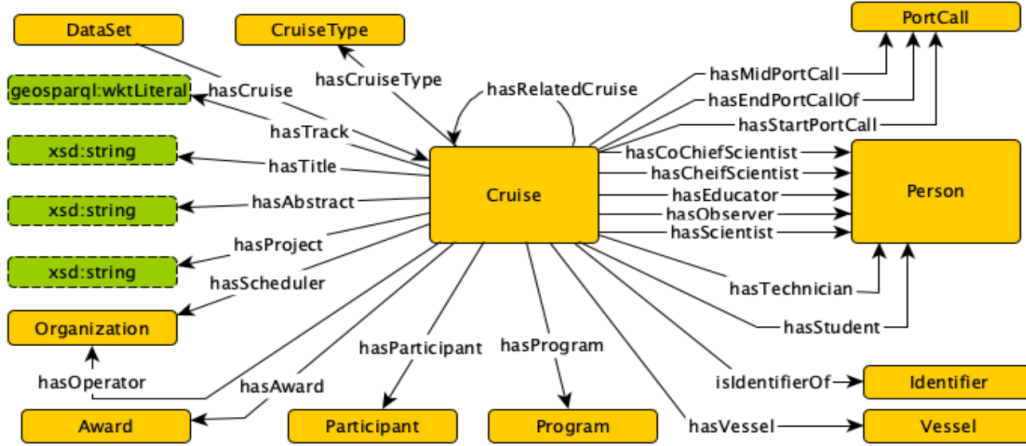


Figure 3.1: *GBO Cruise Ontology Schema*

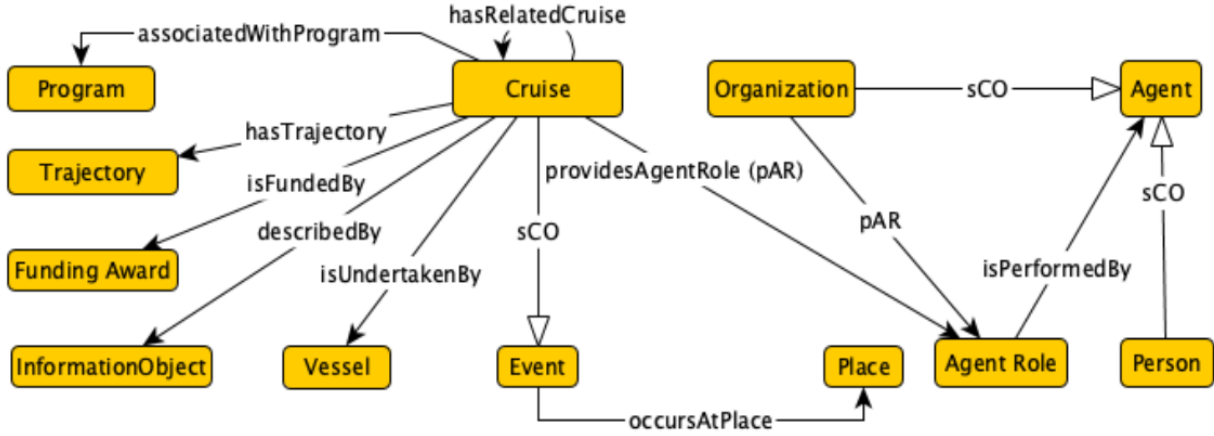


Figure 3.2: *GMO Cruise Ontology Schema*, (*sCO*=*subClassOf* and *pAR*=*providesAgentRole*)

Cruise class in the GeoLink Base Ontology (GBO) and the GeoLink Modular Ontology (GMO), respectively. The Cruise class is defined with a rich set of both object and data properties, enabling a semantically expressive representation. Certain data properties—such as *hasTitle*, *hasProject*, and *hasAbstract*—offer straightforward string-based comparison for instance matching tasks. However, the ontological structure also supports more fine-grained alignment by allowing traversal across interconnected object and data properties, facilitating deeper semantic comparisons.

A key strength of the GeoLink Cruise Benchmark lies in its inclusion of a wide range of

OWL and OWL2 axioms, which allow for a more comprehensive evaluation of ontology-aware instance matching systems. The ontology makes use of advanced OWL vocabulary such as `owl:sameAs`, `owl:priorVersion`, `owl:versionInfo`, `owl:imports`, `owl:unionOf`, `owl:inverseOf`, and `owl:complementOf`. These constructs enable reasoning beyond basic structural or label similarity and support systems that leverage logical inference and ontology reasoning mechanisms.

Unlike synthetic benchmarks, which typically generate test cases by introducing surface-level changes (e.g., typographical errors or minor lexical variations), the GeoLink Cruise Benchmark enables evaluation beyond label-based matching. Its real-world complexity means that entity matches cannot always be identified through superficial comparisons of class labels or data property values. Instead, successful coreference resolution may require deeper exploration of the graph, including the traversal of object properties and the examination of semantically rich linked data.

For example, consider two Cruise instances from BCO-DMO and R2R that are actually coreferent. While these instances may lack identical labels, they share object properties like `hasProgram`, which link to additional nodes containing properties such as `hasTitle`, `hasAcronym`, `description`, `name`, and `label`. These properties provide informative string values that can be compared using similarity metrics. However, only systems capable of deep graph exploration and contextual reasoning are likely to detect such correspondences.

In summary, the GeoLink Cruise Benchmark challenges coreference resolution systems to go beyond surface-level data matching and to engage with the structural and semantic richness of real-world linked data. This makes it particularly valuable for evaluating system performance in realistic scenarios, where the identification of equivalent instances often depends on more than just label or data property similarity.

Chapter 4

Complex Ontology Alignment Using Large Language Models

4.1 Overview

Most of the work on ontology alignment is evaluated by the OAEI, an annual event focusing on ontology-matching systems' performance, starting in 2004, with focus mostly on simple alignment. A complex alignment track was started in 2018 because simple matching is often not sufficient to capture the rich semantics needed for advanced applications using realistic ontologies. In the context of this track, the datasets utilized throughout these years encompass Complex Conference, Populated Complex Conference, Hydrography, GeoLink, Populated GeoLink, Populated Enslaved, and Taxon. It should be noted, however, that some datasets have ceased to be evaluated in recent years as each track depends on volunteers to run it.

Ontologies that exhibit high conceptual clarity are inherently more reusable, as they are easier to comprehend and interpret. Consequently, a central goal in ontology engineering is the development of modeling methodologies that promote clarity, reusability, and maintainability. It is widely accepted that adherence to established modeling principles enhances the overall quality and usability of ontologies. Reusing ontologies in new contexts often presents

significant challenges despite these best practices. These challenges include (i) discrepancies in granularity between existing ontologies and their intended use cases, (ii) ambiguous or poorly defined concepts, (iii) the inherent complexity of ontology alignment, and (iv) a lack of dedicated tools to support ontology reuse. To address these issues, the Modular Ontology Modeling (MOMo) approach was introduced. MOMo advocates for the creation of small, coherent, and independently understandable ontology modules, each encapsulating a principal domain concept.

A key strength of MOMo is its support for modularity, where each module can be developed, maintained, and reused independently. This modular structure enables traceability through comprehensive documentation and metadata, making it easier to link ontology components to their original design requirements. Additionally, modularization simplifies maintenance by isolating the impact of changes and enhances adaptability by allowing modules to be replaced or refined with minimal disruption to the overall ontology.

In the context of the Semantic Web, the term module can have various interpretations. In this work, it specifically refers to a self-contained portion of an ontology that captures a key concept along with its essential features. For example, a Person module may include related elements such as BirthEvent, PersonalInfoItem, and Credential. Modules fulfill a dual role: They serve as technical constructs for grouping semantically related classes and properties while also reflecting the conceptual structure understood by domain experts. Despite potential overlaps or hierarchical relationships, this modular architecture supports an intuitive and expert-aligned representation of knowledge.

Modularization facilitates a systematic approach to ontology construction, allowing for the development and evaluation of individual components before integrating them into a larger framework. This not only supports the scalability and coherence of large ontologies but also improves accessibility for both developers and domain experts. Furthermore, the ability to interchange modules based on context or level of detail fosters reusability and adaptability.

In this study, we investigate the role of incorporating module-level information in improving the identification of complex ontology alignments. Specifically, we leverage the

GeoLink Modular Ontology (GMO)—a schema designed as part of the GeoLink project to integrate several large-scale ocean science data repositories. We enrich alignment tasks with module-level metadata, including descriptions, core axioms (when applicable), and existing alignment references as specified in the GeoLink Core Ontology Design Patterns. As demonstrated in the following sections, this modular context is critical for resolving complex alignment scenarios where semantic correspondences cannot be determined solely through label or structure-based similarity.

Adapting large language models (LLMs) for specific tasks can be accomplished through two primary strategies: fine-tuning and prompt engineering. Fine-tuning involves retraining the model on domain-specific data, which, while effective, demands substantial computational resources, technical expertise, and time. Moreover, it carries the risk of overfitting, potentially reducing the model’s generalizability to other tasks or domains.

In contrast, prompt-based adaptation offers a more resource-efficient alternative. Prompting leverages a pre-trained LLM without modifying its underlying parameters, enabling rapid and flexible task adaptation. This approach requires minimal computational power and can be deployed effectively by users without deep machine learning expertise. As such, prompting has become increasingly popular as a practical and scalable solution for customizing LLM behavior.

Prompt engineering plays a critical role in optimizing LLM performance across diverse tasks. While models can generate outputs using zero-shot prompting—responding to task instructions without seeing examples—more complex tasks often require advanced prompting techniques to guide the model’s reasoning and improve response quality. Among these, the inclusion of contextual information and structured reasoning processes has shown substantial performance gains.

Several task-agnostic prompting techniques have emerged as effective tools in this regard:

- Zero-shot prompting: The model receives a task description in natural language at inference time, without any training examples.
- Few-shot prompting: The model is provided with a limited number of task-specific

examples that demonstrate the desired input-output behavior, improving its contextual understanding.

- Chain-of-thought prompting: This method introduces intermediate reasoning steps in the prompt, helping the model decompose and solve complex tasks through structured logical progression. Collectively, these techniques demonstrate that careful prompt design can significantly enhance LLM effectiveness, offering a practical alternative to model fine-tuning for task specialization.

These techniques enable large language models to adapt and respond to a wide variety of tasks with varying levels of guidance and specificity. Figure 4.1 illustrates our workflow, which begins by uploading the entire GMO file as an initial prompt. Subsequently, we include specific entities from the second ontology, the GBO, to inquire about their alignment with the GMO. If the initial prompt successfully provides the relevant segments, further prompts in the chain-of-thought process are unnecessary. Otherwise, we can provide a list of all GMO module names and ask GPT to identify the most related modules for the GBO entities. Then, in a subsequent prompt, we supplement the inquiry with module information and request the related segments again.

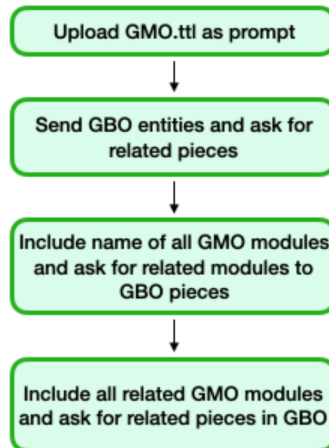


Figure 4.1: *Prompting workflow*

4.2 Contribution

This work investigates the potential of large language models (LLMs), specifically OpenAI’s GPT-4, for performing complex ontology alignment tasks by leveraging modular ontology information⁵. The study involves uploading the GeoLink Modular Ontology (GMO) in RDF Turtle format to GPT-4, which is capable of parsing the data and responding to alignment queries, albeit with some latency. The goal is to assess whether GPT-4 can identify complex alignment rules expressed in logical form, such as those linking GBO entities (e.g., `hasCoPrincipalInvestigator` and `Award`) to their semantically corresponding entities in GMO.

To support this process, the GBO entities involved in the alignment rule were extracted directly from the ontology in their original RDF format. Importantly, the approach relies on the inclusion of module-level metadata, based on the argument that flat ontologies lacking internal structure often fail to provide sufficient semantic cues for automated alignment. The study emphasizes that embedding module information—either during ontology development or through documentation—can significantly reduce ambiguity and improve the performance of ontology alignment systems.

Initial findings indicate that GPT-4, when guided with relevant module context, can generate promising alignment suggestions. However, further work is required to formalize this into an autonomous alignment system with high precision. The authors propose a human-in-the-loop approach as an intermediate step, where ontology engineers verify and refine LLM-generated suggestions. This strategy not only reduces the human effort required for alignment but also guides the LLM toward more accurate and complete outputs, supporting the broader vision of LLM-assisted semantic integration.

This study also reinforces the need to improve ontology modeling methodologies to include modular design as a standard practice, thereby enhancing the reusability, maintainability, and alignability of ontologies.

The performance of GPT-4 in detecting complex ontology alignments was assessed using standard evaluation metrics—recall and precision. In one illustrative example, GPT-4 correctly identified all four expected entities within the GeoLink Modular Ontology (GMO),

without introducing any incorrect entities. This demonstrates a case of both high recall (indicating complete coverage of expected entities) and high precision (indicating no inclusion of irrelevant entities).

Recall, in this context, reflects the system’s ability to retrieve all relevant entities involved in a complex alignment rule. The analysis shows that for a substantial portion of the evaluated rules, GPT-4 was able to identify the majority of the expected GMO entities. In many instances, the system achieved full recall, demonstrating its effectiveness when guided by appropriate prompts and module-level information.

Precision measures the accuracy of the entities returned—specifically, the extent to which the suggestions made by the model align with the correct target entities. The results show that, in most cases, the responses were highly precise, with many instances achieving full precision. This implies that the model rarely suggested irrelevant entities, which is crucial for reducing manual verification effort in downstream tasks.

Moreover, the integration of module information significantly contributed to both higher recall and precision, highlighting the importance of structured ontological context in improving alignment quality. Overall, the evaluation indicates that GPT-4, when supported with appropriate prompts and modular schema information, is capable of producing accurate and comprehensive alignment suggestions with minimal noise.

Chapter 5

Conclusion

5.1 Summary

This work presents a significant advancement in the domain of coreference and complex ontology alignment by introducing both a realistic benchmark and a neural-symbolic alignment system. First, we proposed the GeoLink Cruise benchmark as a novel and realistic evaluation dataset for coreference resolution. Unlike synthetic datasets, our benchmark, accepted by OAEI 2020, captures the intricacies of real-world data, particularly by leveraging semantically linked and manually curated geospatial and temporal entity equivalences.

Secondly, we introduced a neural-symbolic system for complex ontology alignment that operates effectively without relying on shared individuals—a limitation that has impeded prior systems. Our approach integrates large language models (LLMs) with symbolic inputs in the form of ontology modules, yielding promising results in both precision and recall, particularly under realistic conditions where ontology structures are ambiguous or heterogeneous.

Finally, our empirical analysis and system implementation demonstrate a promising path forward in automating complex alignment tasks. We emphasize the critical role of module information in improving alignment quality and highlight the broader implications for ontology engineering practices, encouraging the inclusion of module structure within ontological

documentation and design.

5.2 Conclusion

Looking ahead, several promising research directions stem from our contributions. We intend to actively maintain and expand the GeoLink Cruise benchmark to ensure its continued relevance for the evaluation of coreference resolution systems, particularly within the OAEI framework. Further assessment of existing systems on this benchmark will provide insights into their capabilities and limitations under realistic conditions. In parallel, we aim to deepen our investigation into the integration of large language models with symbolic representations by exploring alternative ways of presenting module information—such as through structured logical axioms—to enhance model performance. Another line of inquiry involves fine-tuning language models specifically for ontology alignment tasks to better capture domain-specific nuances. We also plan to develop semi-automated, human-in-the-loop approaches, where ontology engineers review and refine model-generated suggestions, thereby balancing efficiency with expert oversight. Additionally, we envision hybrid systems that combine traditional alignment algorithms with LLM-based reasoning to tackle complex alignment scenarios more effectively. Collectively, these efforts aim to advance the development of ontology alignment systems that are not only accurate and scalable but also practical for deployment in real-world semantic data environments.

Bibliography

- [1] Richard Cyganiak. The linking open data cloud diagram. <http://richard.cyganiak.de/2007/10/lod/>, 2011.
- [2] Reihaneh Amini, Michelle Cheatham, Pawel Grzebala, and Helena B. McCurdy. Towards best practices for crowdsourcing ontology alignment benchmarks. In *Proceedings of the 11th International Workshop on Ontology Matching (OM-2016)*, pages 1–12, 2016.
- [3] Reihaneh Amini, Lu Zhou, and Pascal Hitzler. Geolink cruises: a non-synthetic benchmark for co-reference resolution on knowledge graphs. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2959–2966, 2020.
- [4] Michelle Cheatham, Adila Krisnadhi, Reihaneh Amini, Pascal Hitzler, Krzysztof Janowicz, Adam Shepherd, Tom Narock, Matt Jones, and Peng Ji. The geolink knowledge graph. *Big Earth Data*, 2(2):131–143, 2018.
- [5] Reihaneh Amini, Sanaz Saki Norouzi, Pascal Hitzler, and Reza Amini. Towards complex ontology alignment using large language models. In *International Knowledge Graph and Semantic Web Conference*, pages 17–31. Springer, 2024.
- [6] T Berners Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):34–43, 2001.
- [7] Daniel L Costa, Michael J Albrethsen, Matthew L Collins, Samuel J Perl, George J Silowash, and Derrick L Spooner. An insider threat indicator ontology. *CERT Center and Carnegie Mellon University Software Engineering Institute* (http://resources.sei.cmu.edu/asset_files/TechnicalReport/2016-005-001_454627.pdf), 2016.

- [8] Pavel Shvaiko and Jérôme Euzenat. Ontology matching: state of the art and future challenges. *IEEE Transactions on knowledge and data engineering*, 25(1):158–176, 2011.
- [9] Lu Zhou, Elodie Thiéblin, Michelle Cheatham, Daniel Faria, Catia Pesquita, Cassia Trojahn, and Ondřej Zamazal. Towards evaluating complex ontology alignments. *The Knowledge Engineering Review*, 35:e21, 2020.
- [10] Michelle Cheatham and Pascal Hitzler. String similarity metrics for ontology alignment. In *The Semantic Web–ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II 12*, pages 294–309. Springer, 2013.
- [11] Tzanina Saveta, Evangelia Daskalaki, Giorgos Flouris, Irimi Fundulaki, Melanie Herschel, and Axel-Cyrille Ngonga Ngomo. Pushing the limits of instance matching systems: A semantics-aware benchmark for linked data. In *Proceedings of the 24th International Conference on World Wide Web*, pages 105–106, 2015.

Appendix A

Contributions

In the following pages, all contributions made for this dissertation are listed in order of appearance in this document.

Towards Best Practices for Crowdsourcing Ontology Alignment Benchmarks

Reihaneh Amini, Michelle Cheatham, Pawel Grzebala, and Helena B. McCurdy

Data Semantics Laboratory, Wright State University, Dayton, Ohio.
{amini.2, michelle.cheatham, grzebala.2, mccurdy.18}@wright.edu

Abstract. Ontology alignment systems establish the links between ontologies that enable knowledge from various sources and domains to be used by applications in many different ways. Unfortunately, these systems are not perfect. Currently the results of even the best-performing alignment systems need to be manually verified in order to be fully trusted. Ontology alignment researchers have turned to crowdsourcing platforms such as Amazon’s Mechanical Turk to accomplish this. However, there has been little systematic analysis of the accuracy of crowdsourcing for alignment verification and the establishment of best practices. In this work, we analyze the impact of the presentation of the context of potential matches and the way in which the question is presented to workers on the accuracy of crowdsourcing for alignment verification.

Keywords: Ontology Alignment, Crowdsourcing, Mechanical Turk

1 Introduction

While the amount of linked data on the Semantic Web has grown dramatically, links between different datasets have unfortunately not grown at the same rate, and data is less useful without context. Links between related things, particularly related things from different datasets, are what enable applications to move beyond individual silos of data towards synthesizing information from a variety of data sources. The goal of ontology alignment is to establish these links by determining when an entity in one ontology is semantically related to an entity in another ontology (for a comprehensive discussion of ontology alignment see [4]).

The performance of automated alignment systems is becoming quite good for certain types of mapping tasks; however, no existing system generates alignments that are completely correct [13]. As a result, there is significant ongoing research on alignment systems that allow users to contribute their knowledge and expertise to the mapping process. Interactive alignment systems exist on a spectrum ranging from entirely manual approaches to semi-automated techniques that ask humans to chime in only when the automated system is unable to make a definitive decision [10]. Because manual alignment is feasible only for small datasets, most research in this area focuses on semi-automated approaches that interact with the user only intermittently. The simplest approach is to send all or a subset of the matches produced through automated techniques to a

user for verification [5]. Other systems only ask the user for guidance at critical decision points during the mapping process, and then attempt to leverage this human-supplied knowledge to improve the scope and quality of the alignment [3].

The issue with the above methods is that ontology engineers and domain experts are very busy people, and they may not have time to devote to manual or semi-automated data integration projects. As a result, some ontology alignment researchers have turned to generic large-scale crowdsourcing platforms, such as Amazon’s Mechanical Turk. Although the use of such crowdsourcing platforms to facilitate scalable ontology alignment is becoming quite common, there is some well-founded skepticism regarding the trustworthiness of crowd-sourced alignment benchmarks. In this work we depart from existing efforts to improve performance of crowdsourcing alignment approaches (e.g. minimizing time and cost) and instead explore whether or not *design* choices made when employing crowdsourcing have a strong effect on the matching results. In particular, there is concern that the results may be sensitive to how the question is asked. The specific questions we seek to answer in this work are:

- Q1: Does providing options beyond simple yes or no regarding the existence of a relationship between two entities improve worker accuracy?
- Q2: What is the impact of question type (e.g. true/false versus multiple choice) on workers’ accuracy?
- Q3: What is the best way to present workers with the contextual information they need to make accurate decisions?
- Q4: It is possible to detect scammers who produce inaccurate results on ontology alignment microtasks?

These are all important questions that must be addressed if researchers in the ontology alignment field are going to accept work on ontology alignments evaluated via crowdsourcing or a crowdsourced alignment benchmark as valid. Section 2 of this paper discusses previous research on crowdsourcing in semi-automated ontology alignment systems. In Section 3, we describe our experimental setup and methodology, and in Section 4 we evaluate the results of those experiments with respect to the research questions presented above. Section 5 summarizes the results and discusses plans for future work on this topic.

2 Background and Related Work

We leverage Amazon’s Mechanical Turk platform in this work. Amazon publicly released Mechanical Turk in 2005. It is based on the idea that some types of tasks that are currently very difficult for machines to solve but are straightforward for humans. The platform provides a way to submit these types of problems, called Human Interface Tasks (HITs), to thousands of people at once. Anyone with a Mechanical Turk account can solve these tasks. People (called Requesters) who send their tasks to Amazon’s servers compensate the people (called Workers or Turkers) who work on the tasks with a small amount of money. Requesters can require that workers have certain qualifications in order to work on their tasks.

For example, workers can be required to be from a certain geographical area, to have performed well on a certain number of HITs previously, or to have passed a qualification test designed by the requester¹.

The primary goal of this work is not to create a crowdsourcing-based ontology alignment system, but rather to begin to determine best practices related to how the crowdsourcing component of such a system should be configured for best results. There has been relatively little research into this topic thus far – most existing work focuses on evaluating the overall performance of a crowdsourcing-based alignment system. An example is CrowdMap, developed in 2012 by Sarasua, Simperl and Noy. This work indicates that working on validation tasks (determining whether or not a given relationship between two entities holds) or identification tasks (finding relationships between entities) are both feasible for workers [12]. Our own previous work has used crowdsourcing to verify existing alignment benchmarks [1] and evaluate the results of an automated alignment system on matching tasks for which no reference alignments are available [2].

The majority of work related to presenting matching questions via a crowdsourcing platform has been done by Mortensen and his colleagues [6–8]. It focused on using crowdsourcing to assess the validity of relationships between entities in a single (biomedical) ontology rather than on aligning two different ontologies, but these goals have much in common. Mortensen noted that in some cases workers who passed qualification tests in order to be eligible to work on the rest of their ontology validation tasks were not necessarily the most accurate, as some of them seemed to rely on their intuition rather than the provided definitions. This led the researchers to try providing the definition of the concepts involved in a potential relationship, which increased the accuracy of workers. The results also indicate that phrasing questions in a positive manner led to better results on the part of workers, e.g. asking whether “A computer is a kind of machine” produced better results than asking whether “Not every computer is a machine.”

Our own work on crowdsourcing ontology alignment and the work of Mortensen describe somewhat ad hoc approaches to finding appropriate question presentation formats and screening policies for workers in order to achieve good results. The work presented here differs from previous efforts by conducting a systematic review of a range of options in an attempt to identify some best practices.

3 Experiment Design

This section describes the experimental setup, datasets, and Mechanical Turk configuration in enough detail for other researchers to replicate these results. The code used is available from <https://github.com/prl-dase-wsu/Ontology-Alignment-mTurk>. The ontologies and reference alignments are from the Conference track of the Ontology Alignment Evaluation Initiative (OAEI).²

¹ <http://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanicalTurkRequester>

² <http://oei.ontologymatching.org/2015/>

3.1 Potential Matches

In order to evaluate the effect of question type, format, and other parameters on worker accuracy, we established a set of 20 potential matches that workers were asked to verify. These matches are all 1-to-1 equivalence relations between pairs of entities drawn from ontologies within the Conference track of the OAEI. Ten of the 20 potential matches are valid. These were taken from the reference alignments. The remaining ten potential matches are invalid. These were chosen based on the most common mistakes within the alignments produced by the 15 alignment systems from the OAEI that performed better than the baseline. For both the valid and invalid matches, we balanced the number of matches in which the entity labels had high string similarity (e.g. “Topic” and “Research.Topic”) and low string similarity (e.g. “Paper” and “Contribution”).

Even though all relations are equivalence, some of our tests offered workers a choice of subsumption relationships. Unfortunately, a primary hindrance to ontology alignment research is the lack of any widely accepted benchmark involving more than 1-to-1 equivalence relations. Until such a benchmark is available, we have limited options. However, the main idea behind our approach here was to provide users with more than a yes-or-no choice. This, together with the precision-oriented and recall-oriented interpretation of responses³, allows researchers to mitigate some of the impacts between people who only answer “yes” in clear-cut cases and those who answer “yes” unless it is obviously not the case.

3.2 Experiment Dimensions

Researchers in this area are so familiar with ontologies and ontology alignment that they risk presenting crowdsourcing workers with questions in a form that makes sense to them but is unintuitive to the uninitiated. We therefore selected the following common methods of alignment presentation for evaluation.

Factor 1: Question Type Previous work has used two different approaches to asking about the relationship between two entities: true/false, in which a person is asked if two entities are equivalent [9], and multiple choice questions, in which the person is asked about the precise relationship between two entities, such as equivalence, subsumption, or no relation [2, 12].

A typical true/false question is “Can Paper be matched with Contribution”? Workers can then simply answer “Yes” or “No.” A multiple choice question regarding the same two entities takes the form “What is the relationship between Paper and Contribution?” and has four possible answers: “Paper and Contribution are the same,” “Any thing that is a Paper is also a Contribution, but anything that is a Contribution is not necessarily a Paper,” “Any thing that is a Contribution is also a Paper, but anything that is a Paper is not necessarily a Contribution” and “There is no relationship between Paper and Contribution.” The motivation for the second of these approaches is that as automated alignment systems attempt to move beyond finding 1-to-1 equivalence relationships

³ These evaluation metrics will be discussed in details in Section 4.1 .

towards identifying subsumption relations and more complex mappings involving multiple entities from both ontologies, the ability to accurately crowdsource information about these more complex relationships becomes more important. Additionally, a common approach taken by many current alignment systems is to identify a pool of potential matches for each entity in an ontology and then employ more computationally intensive similarity comparisons to determine which, if any, of those potential matches are valid. If crowdsourcing were to be used in this manner for semi-automated ontology alignment, one approach might be to use the multiple choice question type to cast a wide net regarding related entities, and then feed those into the automated component of the system.

Factor 2: Question Format A primary purpose of ontologies is to contextualize entities within a domain. Therefore, context is very important when deciding whether or not two entities are related. Even in cases where the entities have the same name or label, they may not be used in the same way. These situations are very challenging for current alignment systems [1]. Providing context is particularly important in crowdsourcing, because workers are not domain experts and so may need some additional information about the entities in order to understand the relation between them. For this reason, we explored the impact of providing workers with four different types of contextual information:

Label Only entity labels (no context) is provided.

Definition A definition of each entity’s label is provided. Definitions were obtained from Wiktionary.⁴ If a label had multiple definitions, the one most related to conferences (the domain of the ontologies) was manually selected.⁵

Relationships (Textual) The worker is presented with a textual description of all of the super class, sub class, super property, sub property, domain and range relationships involving the entities. The axioms specifying these relations were extracted from the ontologies and “translated” using Open University’s OWL to English tool.⁶ An example for “Evaluated_Paper” is:

- *No camera ready paper is an evaluated paper.*
- *An accepted paper is an evaluated paper.*
- *A rejected paper is an evaluated paper.*
- *An evaluated paper is an assigned paper.*

Relationships (Graphical) The worker is presented with the same information as above, but as a graph rather than as text. The relationships involving both entities from the potential match are shown in the same graph, with an edge labeled “equivalent?” between the entities in question. Figure 1 shows an example for “Place.”

⁴ https://en.wiktionary.org/wiki/Wiktionary:Main_Page

⁵ Note that the goal of this work is to determine the best way in which to prevent matching-related questions rather than to create a fully automated approach; however, the step of choosing the most relevant definition of a label could be automated in future work.

⁶ <http://swat.open.ac.uk/tools>

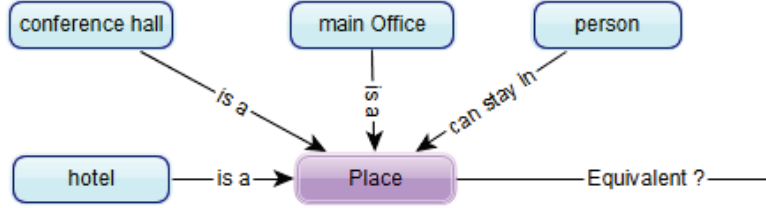


Fig. 1: Graphical depiction of the relationships involving the entity “Place”

3.3 Mechanical Turk Setup

We tested all combinations of question type and format described above, for a total of 8 treatment groups. HITs for each of these tests contained the 20 questions described in Section 3.1. 160 workers were divided among the treatment groups. They were paid 20 cents to complete the task.

One important missing point in current related work is whether workers were prevented from participating in more than one treatment group of the experiment, a potential source of bias. For example, if workers participate in the

Task

Read the definitions of **Topic**, **Research_Topic**, and then select one of the choices below:

Label	Definition
Topic:	The subject of a workshop or session at a conference
Research_topic:	The subject of a paper or discussion regarding some form of research

Based on the above definitions, what is the relation between **Topic** and **Research_Topic**?

- ☐ "Topic" and "Research_Topic" mean the same thing.
- ☐ Any thing that is "Topic" is also "Research_Topic", But anything that is "Research_Topic" is NOT necessarily "Topic".
- ☐ Any thing that is "Research_Topic" is also "Topic", But anything that is "Topic" is NOT necessarily "Research_Topic".
- ☐ There is no relation between "Topic" and "Research_Topic".

Fig. 2: An example of multiple choice HIT containing entity definitions on Amazon’s Mechanical Turk server

definitions treatment group and then work on the graphical relationships tasks, they may remember some definitions and that may influence their answers. In order to avoid this source of bias, we created a Mechanical Turk qualification, assigned this to any worker who completed one of our HITs and specified that our HITs were only available to workers who did *not* possess this qualification.

Finding capable and diligent workers is always a difficult problem when using a crowdsourcing platform. One common approach is to require a worker to pass a qualification test before they are allowed to work on the actual tasks. Although this strategy seems quite reasonable, qualification tasks are generally short and

contain only basic questions, so a worker’s performance on it is not always reflective of their performance on the actual tasks. Furthermore, sometimes workers will take the qualification task very seriously but then not apply the same level of diligence to the actual tasks. Additionally, workers tend to expect to be compensated more if they had to pass a qualification test. Another approach to attracting good workers is to offer a bonus for good performance [14]. Many requesters also use “candy questions” that have an obviously correct answer, in order to detect bots or people who have just randomly clicked answers without reading the questions. Requesters generally ignore the entire submission of any worker who misses a candy question. We have employed all of these strategies in the course of this work. The results we obtained from workers who passed a qualification test containing simple questions of the type we intended to study were not encouraging – we qualified workers who achieved greater than 80% accuracy on a qualification test; however, those workers delivered poor performance on the actual tasks (average accuracy 51%). As mentioned previously, other researchers experienced a similar problem [11]. As a result, we decided against using qualification tests and settled on offering workers a \$2 bonus if they answered 80% or more of the questions correctly. Of course, this particular strategy is only applicable in situations in which the correct answers to the questions are known in advance. In the future, we plan to more systematically explore the ramifications of different methods for dealing with unqualified, unethical, and lazy workers.

4 Analysis of Results

4.1 Impact of Question Type

Ontology alignments are typically evaluated based on precision (how many of the answers given by a person or system are correct) and recall (how many of the correct answers were given by a person or system). These metrics are based on the number of true positives, false positives and false negatives. The meaning for this is clear when we are discussing 1-to-1 equivalence relations (i.e. in the true/false case) but it is less obvious how to classify each result in the multiple choice case, where subsumption relations are possible. For example, consider the multiple choice question in Figure 2. According to the reference alignment, “Topic” and “Research_Topic” are equivalent. It is therefore clear that if the user selects the first multiple choice option, it should be classified as a true positive, whereas selecting the last option should count as a false negative. But how should the middle two options be classified? Unfortunately, most previous work that allows users to specify either equivalence or subsumption relations is vague about how this is handled [12].

In this work we take two different approaches to classifying results as true positives, false positives, or false negatives. In what we call a **recall-oriented** analysis, we consider a subsumption answer to be effectively the same as an equivalence (i.e. identification of *any* relationship between the entities is considered as agreement with the potential match). In the example above, this would result in the middle two options being considered true positives. This approach

allows us to evaluate how accurate workers are at separating pairs of entities that are related in some way from those that are not related at all. This capability is useful in alignments systems to avoid finding only obvious matches – entities related in a variety of ways to a particular entity can be gathered first and then further processing can filter the set down to only equivalence relations. The other approach, which we call a **precision-oriented** analysis, a subsumption relationship is considered distinct from equivalence (i.e. a potential match is only considered validated by a user if they explicitly state that the two entities are equivalent). This would result in options two and three from the example above being classified as false negatives. This interpretation may be useful for evaluating an alignment system that is attempting to find high-quality equivalence relations between entities, which it may subsequently use as a seed for further processing.

The overall results based on question type provided in Figure 3 show that workers have more balanced precision and recall on True/False questions than on Multiple Choice ones. While this is intuitive [2], it is helpful to have quantitative data for the different question types on the same set of potential matches. Also, some interesting observations can be made based on these results, including:

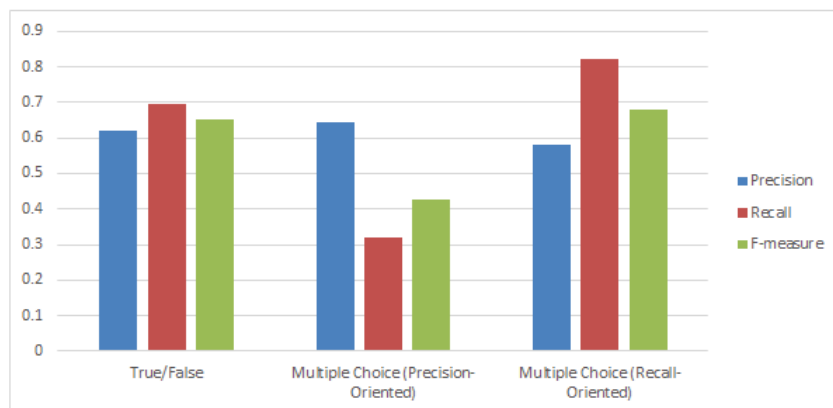


Fig. 3: Workers’ performance on true/false and multiple choice questions

Workers are relatively adept at recognizing when *some* type of relationship exists between two entities. The F-measure of 0.65 on the true/false questions and 0.67 using the recall-oriented analysis of the multiple choice questions tells us that workers can fairly accurately distinguish the entities that are somehow related to each other from those that are not, regardless of the question type used to solicit this information from them. In fact, the multiple choice type of question resulted in significantly higher recall (0.82 versus 0.69

for true/false), making it an enticing option for ontology alignment researchers interested in collecting a somewhat comprehensive set of potential matches.

Workers appear to perform poorly at identifying the *type* of relationship that exists between two entities. This claim is less strong than the previous one, because according to our reference alignments, the only relationship that ever held between entities was equivalence. Unfortunately, there are no currently accepted alignment benchmarks that contain subsumption relations, so confirmation of these results is a subject for future work. However, the F-measure of the precision-oriented analysis of the multiple choice questions (0.42, as shown in Figure 3) clearly indicates that the workers did not do well at classifying nuanced relationships between entities.

If precision is paramount, it is best to use true/false questions. While the precision-oriented analysis of the multiple choice questions results is very slightly higher precision than the true/false questions (0.62 versus 0.64), its recall is so low as to be unusable (0.32). If ontology alignment researchers wish to validate 1-to-1 equivalence relationships generated by their system or establish high-quality “anchor” mappings that can be used to seed an alignment algorithm, we recommend that they present their queries to workers as true/false questions.

4.2 Impact of Question Format

As shown in Figure 4, there is a fairly wide range in F-measure for the four question formats, 0.54 to 0.67. Within a single question type, for example true/false, the F-measure varies from 0.59 when no context is provided to 0.73 when workers are provided with the definitions of both terms. This is somewhat surprising, since the domain covered by these ontologies is not particularly esoteric or likely to contain many labels that people are not already familiar with. We note the following observations related to this experiment.

Workers leverage contextual information when it is provided, and this improves their accuracy. Other researchers have speculated that workers may rely on their intuition more than the provided information to complete this type of task, but that hypothesis is not supported by the results here – there is a distinct difference in precision, recall, and F-measure when workers have some contextual information than when they are forced to decide without any context.

When precision is important, providing workers with definitions is effective. The previous section indicated that when the task is to accurately identify equivalent entities, the True-False question style is the best approach. Now Figure 4 indicates that the best accuracy in this situation occurs when workers are provided with entity definitions (F-measure 0.73), while the worst case is when workers are given a piece of the ontology’s schema or just the entities’ names (F-measure 0.61 and 0.58, respectively).

When finding entity pairs that have *any* relationship is the goal, a graphical depiction is helpful. The recall-oriented analysis of multiple choice questions showed relatively high recall and F-measure for all question formats, with recall of the graphical format slightly edging out that of label definitions. Furthermore, by calculating the True Negative Rate (TNR) of these different formats for multiple choice questions, we discovered that when provided with a graphical depiction of entity relationships, workers more accurately identified when the two entities in the potential match were not related at all (TNR 0.70).

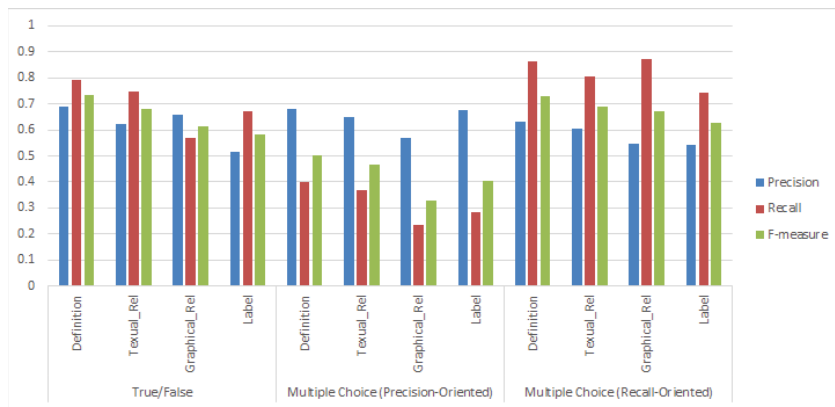


Fig. 4: Workers’ performance based on question format

4.3 Dealing with Scammers

Avoiding or handling scammers (people who try to optimize their earnings per time spent) is a recurring theme in crowdsourcing-related subjects. During the presentation of the authors’ own work related to crowdsourcing in ontology alignment [1], several attendees expressed the notion that time is likely a useful feature with which to recognize scammers. The intuition is that scammers rush through tasks and quickly answer all of the questions without taking the time to understand and consider each one. To test this hypothesis, we examined the relationship between the time workers spent on a HIT and their accuracy across all question types and formats. For this, we used the “Accept” and “Submit” timestamps included with the Mechanical Turk results available from Amazon. Following is a list of our observations based on this data.

Time spent on a task is a poor indicator of accuracy. We first looked at the average time spent on the HIT by high-performing workers (those who answered more than 80% of the questions within the HIT correctly) and low-performing workers (those who answered fewer than half of the questions correctly). The results were unexpected: high-performing workers spent less than five minutes on the task while low-performers averaged seven minutes.

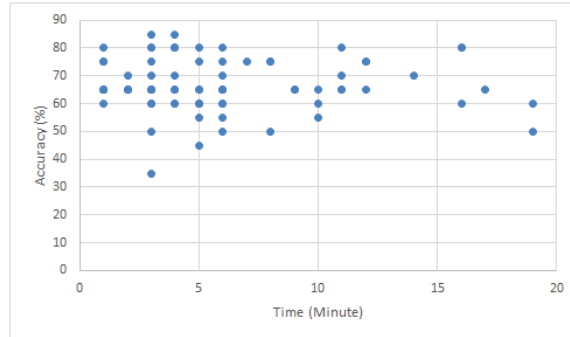


Fig. 5: Average accuracy of workers based on time spent

The above observation holds even at the extreme ends of the time spectrum. Even workers who answered all 20 questions in an extremely short time, such as one or two minutes, did not always have poor accuracy. For instance, multiple workers who spent less than a minute on true/false questions had an accuracy between 60% and 70%, which is close to the overall average on that question type. Conversely, several workers who spent more than 8 minutes had an accuracy between 45% and 55%. It therefore seems that setting thresholds for time to recognize scammers is not a viable strategy.

5 Conclusions and Future Work

The idea of using crowdsourcing for ontology alignment has been gaining in popularity over the past several years. However, very little systematic work has yet gone into how best to present potential matches to users and solicit their responses. This work has begun an effort towards establishing some best practices in this area, by exploring the impact of question type and question format on worker accuracy. Additionally, a popular strategy of mitigating the impact of scammers on accuracy was explored. The results of some experiments confirm common intuition (e.g. workers are better able to determine when any relationship exists between two entities than they are at specifying the precise nature of that relationship), while other results refute popularly held beliefs (e.g. scammers cannot be reliably identified solely by the amount of time they spend on a task). Our overall recommendations are that users interested in verifying the accuracy of an existing alignment or establishing high-quality anchor matches from which to expand are likely to achieve the best results by presenting the definitions of the entity labels and allowing workers to respond with true/false to the question of whether or not an equivalence relationship exists. Conversely, if the alignment researcher is interested in finding entity pairs in which *any* relationship holds, they are better off presenting workers with a graphical depiction of the entity relationships and a set of options about the type of relation that exists, if any.

This work is relevant not only to crowdsourcing approaches to ontology alignment, but also to interactive alignment systems, as well as to user interfaces that attempt to display the rationale behind the matches that make up an alignment generated through other means. However, there are other aspects that are specific to crowdsourcing that should be further explored such as, the best way of enticing large numbers of capable workers to complete alignment tasks in a timely manner. We plan to address this challenge in our future work on this topic.

References

1. Cheatham, M., Hitzler, P.: Conference v2.0: An uncertain version of the OAEI Conference benchmark. In: *Proceedings of the International Semantic Web Conference*, pp. 33–48. Springer (2014)
2. Cheatham, M., Hitzler, P.: The properties of property alignment. In: *Proceedings of the 9th International Conference on Ontology Matching*, vol. 1317, pp. 13–24. CEUR-WS. org (2014)
3. Cruz, I.F., Stroe, C., Palmonari, M.: Interactive user feedback in ontology matching using signature vectors. In: *Proceedings of the International Conference on Data Engineering (ICDE)*, pp. 1321–1324. IEEE (2012)
4. Euzenat, J., Shvaiko, P.: *Ontology Matching*, vol. 333. Springer (2007)
5. Kheder, N., Diallo, G.: Servombi at OAEI 2015. *Proceedings of the 12th International Workshop on Ontology Matching* p. 200 (2015)
6. Mortensen, J., Musen, M.A., Noy, N.F.: Crowdsourcing the verification of relationships in biomedical ontologies. In: *Proceedings of the AMIA Annual Symposium* (2013)
7. Mortensen, J.M.: Crowdsourcing ontology verification. In: *Proceedings of the International Semantic Web Conference*, pp. 448–455. Springer (2013)
8. Mortensen, J.M., Musen, M.A., Noy, N.F.: Ontology quality assurance with the crowd. In: *AAAI Conference on Human Computation and Crowdsourcing* (2013)
9. Noy, N.F., Mortensen, J., Musen, M.A., Alexander, P.R.: Mechanical Turk as an ontology engineer?: Using microtasks as a component of an ontology-engineering workflow. In: *Proceedings of the 5th Annual ACM Web Science Conference*, pp. 262–271. ACM (2013)
10. Noy, N.F., Musen, M.A., et al.: Algorithm and tool for automated ontology merging and alignment. In: *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00)*. Available as SMI technical report SMI-2000-0831 (2000)
11. Oleson, D., Sorokin, A., Laughlin, G.P., Hester, V., Le, J., Biewald, L.: Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. *Human Computation* 11(11) (2011)
12. Sarasua, C., Simperl, E., Noy, N.F.: Crowdmap: Crowdsourcing ontology alignment with microtasks. *Proceedings of the International Semantic Web Conference* pp. 525–541 (2012)
13. Shvaiko, P., Euzenat, J.: Ontology matching: State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering* 25(1), 158–176 (2013)
14. Wang, J., Ghose, A., Ipeirotis, P.: Bonus, disclosure, and choice: What motivates the creation of high-quality paid reviews? (2012)



GeoLink Cruises: A Non-Synthetic Benchmark for Co-Reference Resolution on Knowledge Graphs

Reihaneh Amini
Kansas State University
reihanea@ksu.edu

Lu Zhou
Kansas State University
luzhou@ksu.edu

Pascal Hitzler
Kansas State University
hitzler@ksu.edu

ABSTRACT

Since over a decade coreference resolution systems have been developed in order to find simple 1-to-1 equivalent mapping (sameAs relations) between instances of different linked datasets and knowledge graphs. Comparative evaluations of instance matching systems can inform us about the performance of such systems regarding artificial benchmarks or real-world data challenges. However, the lack of real data for evaluating these systems is currently a bottleneck. In this paper, we propose the use of the Cruise entities in the GeoLink data repository as a real-world instance matching benchmark for linked data and knowledge graphs. The GeoLink project has brought together seven datasets related to geoscience research. Both the ontology (T-box) and the instance data (A-box) of GeoLink are significantly larger than current benchmarks, and they have particularly interesting challenges, such as geospatial and temporal data. The benchmark we propose here consists of two real-world datasets in GeoLink called R2R data and BCO-DMO which includes manual curated owl:sameAs links between more than 900 Cruise entities of these two datasets. The reference alignment was discussed and generated by domain experts from different institutions and is expressed in the Alignment API format.

CCS CONCEPTS

• **Information systems** → **Information integration; Resource Description Framework (RDF); Web Ontology Language (OWL); Ontologies**; • **Computing methodologies** → **Ontology engineering**.

KEYWORDS

Coreference Resolution; Instance Matching; Semantic Web; Benchmarks; Linked Data; Knowledge Graphs

ACM Reference Format:

Reihaneh Amini, Lu Zhou, and Pascal Hitzler. 2020. GeoLink Cruises: A Non-Synthetic Benchmark for Co-Reference Resolution on Knowledge Graphs. In *The 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3340531.3412770>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-6859-9/20/10...\$15.00

<https://doi.org/10.1145/3340531.3412770>

1 INTRODUCTION

A massive amount of data is being piped into the world every second by different sources but there is currently not a consistent and steady approach to regulate the format and forms of this data. Linked data emerged to address challenges in data integration and its goal was to introduce techniques and standards to make data accessible and consumable by both humans and machines. According to Tim Berners-Lee, the Semantic Web is not concerned only with putting data on the web but is also about establishing links between that data in order to put it in context. Who should generate these links, though? Often, related data is published by different people at different times, and the publishers may have less knowledge of one another. As a result, there is a need to create systems capable of identifying meaningful links between datasets in an automated way. This is the goal of coreference resolution (alternatively called instance matching). Coreference resolution systems attempt to determine when two URIs refer to the same individual. Unlike complete ontology alignment systems, which endeavor to find many types of relationships (e.g. equivalence, subsumption) between schema-level entities, coreference resolution systems generally focus on identifying 1-to-1 sameAs relationships between instances.

Because the strong performance of coreference resolution systems is vital to harnessing the full power of linked data, it is important to have a benchmark with real data and real challenges with which to evaluate their performance and to spur innovation in the field. In this paper, we propose the GeoLink Cruises benchmark, based on two different real-world geoscience research datasets that were integrated as part of the NSF GeoLink project. We show that this benchmark is relevant to assessing the performance of existing coreference resolution systems in linked data, while also possessing features with the potential to drive innovation in the field.

The main contributions of this paper are therefore the following:

- Presentation of two ontologies to support semantic- and schema-based evaluations.
- Presenting a benchmark with curated sameAs links between over 900 Cruise instances.
- Publication of the benchmark dataset and alignment under a CC-BY license.

This paper is organized as follows. Section 2 discusses the current coreference resolution benchmarks in the Ontology Alignment Evaluation Initiative (OAEI). In Section 3, we provide information about the GeoLink knowledge graph including some of its specific features. Section 4 discusses the coreference resolution benchmark that we are proposing, followed by some analysis of the performance of current coreference resolution algorithms on our data in Section 5. In Section 6, we conclude with a discussion of potential future work in this area.

2 RELATED WORK

In this section, we provide an overview of current main benchmarks in evaluating coreference resolution systems in linked semantic data and also discuss different algorithms for coreference resolution systems and use that as a basis for enumerating a set of desirable qualities for a benchmark with which to evaluate the performance of such systems.

2.1 Coreference Resolution Benchmarks

There are several existing coreference resolution benchmarks in semantically linked data. The presiding existing benchmark is the Ontology Alignment Evaluation Initiative (OAEI) ¹, which has held an instance matching track since 2009 [8]. The OAEI has been the foremost venue for researchers focused on ontology alignment and coreference resolution to evaluate and showcase their work. The OAEI introduced this track in 2009, and every year since then, a few coreference resolution systems have participated in the track. This track evaluates system performance by comparing the degree of similarity between pairs of instances with a synthetic benchmark called the Semantic Publishing Instance Matching Benchmark (SPIMBENCH) [18]. SPIMBENCH allows for systematic scalability testing, supports a wider range of test cases, and provides an enriched gold standard. The goal of SPIMBENCH track in OAEI is to resolve coreferences to the same Creative Work. The schema of SPIMBENCH contains 22 classes, 85 object properties, and 31 data properties. SPIMBENCH creates datasets by making various transformations to the source file. Value transformations involve syntactic changes such as misspellings, abbreviations, and synonyms. Structural changes involve modifying the depth of the representation, e.g. by nesting or aggregating properties. Semantic variations involve OWL constructs such as introducing equivalence, disjointness, or complex class definitions between schema level entities. So all these modifications to the source data will generate a set of matches and non-matches as synthetic test cases. So SPIMBENCH does not contain any coreferences within itself because of its synthetic type. SPIMBENCH has also two versions: a small-scale “sandbox” and a medium-scale “mainbox”. Because the data is syntactically generated, the scalability is a modifiable factor and if scalability wants to be tested, SPIMBENCH can synthetically generate many more test cases to evaluate the coreference resolution system’s scalability aspect.

SPIMBENCH is very valuable because complete reference alignments allow both precision and recall to be assessed. In addition, the performance of coreference resolution systems in the face of specific types of characteristics, such as semantic or structural variation, can be isolated and evaluated by the synthetic value tasks. On the other hand, these tasks are of necessity somewhat narrowly focused, and it is not clear how predictive performance on these tasks will be on coreference resolution tasks that differ significantly in scope or domain. In particular, both the source and target dataset use the same ontology (because the datasets contain information from the same domain) which also is not similar to real-world data. Having a semantic-aware aspect is a powerful side of SPIMBENCH benchmark because, in the real semantic world, data includes the RDFS axioms. However, ontologies may include more OWL axioms that

should be included in the benchmarks to evaluate the performance of the systems regarding a wide and unexpected range of these axioms, not only a limited set of them. Examples of such axioms include `owl:topDataProperty`, `owl:bottomDataProperty`, `owl:thing`, `owl:nothing`, `owl:real`, and so on. Although a synthetic benchmark is very helpful for evaluating systems on many narrow areas, it ingests bias into the evaluation because the false test cases are built on assumptions like string similarity between equivalent classes, and real-world cases may, and do, look different. For example, “paint” and “color” are instances of the same class without any string similarity, or in another case, “Organizer” might refer to the class of “Person” in one dataset and the class of “Company” in another dataset.

Other synthetic instance matching benchmarks independent of the OAEI are similar. For instance, ONTOBI generates ontologies with a moderate number of instances from Wikipedia infoboxes. The size of the reference ontology is again small (17 classes, 13 object properties, and 128 data properties) and the source and target datasets cover the identical domain [22].

As concluding the above-mentioned points, having synthetic data is valuable in evaluating the scalability of the systems and also to avoid the messiness of real data in many cases. In particular, having a synthetic dataset is very important for initial research; however, it cannot fully replicate real-world cases and examples. As a result, there should be both synthetic datasets for initial research and real-world datasets for replicating the complexity of the real data and evaluating the performance of IM systems facing the real world.

The GeoLink Cruise benchmark differs in that both its schema and A-box are considerably more complex since the data is real and the two datasets are based on different larger schemas.

2.2 Coreference Resolution Systems

There have been a total of 20 unique instance matching systems in semantic structured data. We have reviewed all of these systems in order to determine the techniques that are common to many different systems. The results of this activity show that current systems can be classified into a few general approaches, as described below.

Direct comparison of all individual pairs: These approaches often perform surprisingly well but are not scalable to large datasets. Examples include STRIM [13], SLINT++ [15], and ObjectCoref [11]. Systems in this category sometimes do some analysis before beginning the matching task to determine what information to use to compare individuals. For example, SLINT+ chooses properties with good coverage and strong discriminating power while ObjectCoref uses `sameAs` and `inverseFunctional` properties to create a training set for learning weightings for property-value pairs.

Comparison of individuals based on a “cloud”: In these approaches, a collection of values is created for each individual, and two individuals are compared using a set-similarity metric over their collections. What is included in the collections varies from system to system? For example, InsMT+ [13] includes the individual’s label and the names of properties specified for that individual. SBUEI [19] includes the individual’s label, its datatype property values, and the datatype values of its direct neighbors. Anchor-Flood

¹<http://oaei.ontologymatching.org/2019/>

[9] is similar to SBUEI except that it includes the property names in addition to their values, and only for the individual in question rather than its neighbors.

Two phase comparison: coarse- and fine-grained: These algorithms avoid expensive comparisons between each individual in the source and target ontologies by using faster but less accurate comparisons to find candidate matches that are then compared using more expensive techniques. Some approaches for the coarse-grained comparisons are finding all target individuals that have a data property value exactly in common with the source individual (EXONA [5] and Serimi [1]), finding all individuals that share a property (RiMOM [23]), and finding all individuals of the same type with a somewhat similar label (AgreementMaker [4]) or an exact match on label or alias (Zhishi.links) [16]. The fine-grained comparisons then find the best match from among all of the candidates based on either label (EXONA, RiMOM, AgreementMaker, and Serimi) or property names and values (Zhishi.links).

Reformulation as a different type of problem: Several systems reformulate the coreference resolution problem as a different type of problem. For example, Lily [21] creates a subgraph for each individual based on an electrical circuit model, CODI [12] transforms the alignment problem into the maximum a posteriori optimization problem using Markov logic, and LN2R [17] treats establishing similarities between individuals as a system of equations.

Based on this review of current systems, we conclude that a coreference resolution benchmark should meet the following criteria:

- A large number of instances, to test scalability and approaches that involve coarse-grained filtering.
- Sufficiently numerous common properties among instances of a given type to support comparison on information beyond instance labels. Sufficiently numerous distinct properties among instances of a given type to test systems that combine schema and instance alignment.
- A rich enough schema to support structural-based approaches.
- Enough sameAs links between instances to support approaches that involve supervised machine learning.

Considering the features in the current benchmark and the points regarding what benchmarks need for properly evaluating these systems, we will discuss how our benchmark is addressing these drawbacks.

3 GEOLINK ONTOLOGY MODELS

The benchmark we propose in this paper is a part of the GeoLink knowledge graph. In this section, we explain what the GeoLink knowledge graph is and how our benchmark emerges from it. GeoLink was a project as part of the NSF EarthCube initiative and included seven diverse geoscience datasets that have been merged into a single data repository. GeoLink's merged ontology is available online², and the merged data is accessible through SPARQL queries³. There are currently 282 classes, 338 properties, 5,118,150 instances, and 45,093,750 triples in the knowledge base.

The overall goal of GeoLink was to merge and unify geoscience data from existing repositories into a single knowledge graph. GeoLink has two different ontologies provided to data providers for publishing their data as RDF triples. These two schemas are the GeoLink Modular Ontology (GMO) and the GeoLink Base Ontology (GBO) [14], which we discuss briefly in this section.

The GeoLink ontologies are developed from ontology design patterns (ODPs) [10] to the high standard of a modular ontology [3, 14]. Each concept like a person or a physical sample that is very frequent in many geoscience repositories was defined as an encapsulated concept (a *module*) in GeoLink in collaboration with a group of domain experts, ontologists, and data providers. All these modules stitched together to form the GMO which enables data providers to publish the part of their data that is related to these concepts, based on GMO vocabulary.

The ODPs represent the concepts within GeoLink that unite the different data repositories. 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 any external vocabulary that best fits the data. Although GMO is following best practices in schema modeling, it is hard to always find the exact relation between your data to some of the concepts designed as ODPs [10]. For example, in the GMO there is an AgentRole class that reifies some relationships. To simplify the mapping, the GBO has been developed, it simplifies the patterns in the GMO [3]. Although the GBO itself does not follow high standards of ontology design, it is more convenient for data providers to publish their data to the GBO. Both ontologies are available online⁴ under a CC-BY License. The two ontologies in the GeoLink project have also been used to establish a complex ontology alignment benchmark, which has been utilized in OAEI 2018 and 2019 [24, 25]. The complex alignment benchmark consists of the GMO and GBO ontologies and captures the complex relationships between these two ontologies.

Many of the data repositories contributing to the GeoLink knowledge graph are funded by government organizations such as the NSF. the GeoLink knowledge graph data repositories are:

- R2R (Rolling Deck to Repository)⁵
- BCO-DMO (Biological and Chemical Oceanography Data Management Office)⁶
- DataONE (Data Observation Network for Earth)⁷
- IEDA (Interdisciplinary Earth Data Alliance)⁸
- IODP (International Ocean Discovery Program)⁹
- LTER (Long Term Ecological Research Network)¹⁰
- MBLWHOI (Marine Biological Laboratory Woods Hole Oceanographic Institution)¹¹

In our proposed benchmark, we only use two of these repositories called R2R and BCO-DMO. These two datasets are the main repositories for oceanographic cruises in GeoLink.

⁴https://figshare.com/articles/Complex_Alignment_Benchmark_GeoLink_Dataset/5907172

⁵<http://www.rvdata.us>

⁶<http://www.bco-dmo.org>

⁷<https://www.dataone.org/>

⁸<http://www.iedadata.org/>

⁹<http://www.iodp.org/>

¹⁰<http://www.lternet.edu/>

¹¹<http://www.mblwhoilibrary.org/>

²<http://schema.geolink.org>

³<http://data.geolink.org>

R2R is the Rolling Deck to Repository, collected by U.S. academic research fleet and includes environmental sensor data. This repository collects data from cruise reports, navigation tracks for cruises, funding awards, vessels, expeditions, and so on.

BCO-DMO is the Biological and Chemical Oceanography Data Management Office (BCO-DMO) which keeps the generated data from oceanographic research and publishes this data online. Much of this data is collected during Cruises for which information is also recorded in R2R.

For generating this GeoLink Cruise benchmark, we considered the part of the GMO and GBO which represent a good and deep enough information about the Cruises, because GeoLink has a big ontology and involving the whole GBO and GMO ontologies may distract our focus on Cruise instances. In fact, we considered the Cruise class since oceanographic cruises are central to the repositories and thus are at the heart of the schema, and most of the other classes, object properties, and data properties are reachable via a few edge traversals from Cruise entities. Figures 1 and 2 provide partial schema diagrams for the GBO and GMO ontologies as relevant for our benchmark.

The **Cruise** pattern is an example of a complex pattern [14] and its notion is central to oceanography data. In both ontologies, there are different classes and we explain some that might require further clarification. **Vessel** represents vessels on which cruises are carried out. **Funding Award/Award** pattern describes the funding awards that fund all kinds of ocean science research activities. A **Program** is a collection of things, including cruises, funding awards, activities, and events. **Event** describes generic events, which may include cruises. **Agent** is a generic class that represents an agent (e.g., a person or an organization) and that agent may perform a role as **AgentRole**. **Place** pattern captures spatial information in the Event pattern above and the rest of the ontology. **Organization** describes organizations, including academic institutions, funding agencies, vessel owners, etc. **Person** appears in a variety of contexts such as Chief Scientist on a cruise, Principal Investigator on a project, a participant in a meeting, or creator of a dataset or paper [2].

4 GEOLINK CRUISE CO-REFERENCE RESOLUTION BENCHMARK

Cruise is one of the three main class types (Cruise, People, Organization) in the GeoLink knowledge base. Therefore, we propose an instance matching benchmark called GeoLink Cruise Benchmark including two Cruise datasets, which are R2R and BCO-DMO (see Section 3). The GeoLink Cruise benchmark has some unique features which distinguish it from other current benchmarks. We explain these features in the following.

Accessibility: The benchmark is publicly available¹² under a CC-BY 4.0 License, which means it is free to download, manipulate, and use for any purposes. Both the GMO and GBO directory in the data include the reference alignment between the ontologies named “refalign.rdf”. The reference alignment is expressed using the Alignment format¹³ by the Alignment API [6], because this is the most prominent syntax for such benchmarks, including SPIMBENCH.

Real-world dataset: The GeoLink project comprises many datasets provided by seven data providers. These datasets have been collected and utilized in the geosciences, which indicates that they are natural, realistic datasets, rather than artificial ones as often used in benchmarks. This is one of the main features of our data which makes it more useful for evaluating the performance of instance matching algorithms for practice. As we discussed earlier in Section 2, assumptions that humans make in regards to generating synthetic data are introducing bias into the data which might lead to biased system evaluation. Having synthetic data is useful for early research and for avoiding the messiness of real data for early system testing. It is also helpful in evaluating the systems in specific areas. However, it cannot sufficiently capture the complexity of real-world data.

Ground Truth: For this benchmark, Cruise instances in BCO-DMO and R2R were compared with each other manually and 491 owl:sameAs links were generated between them by data providers. This large number of curated links can evaluate the precision and recall of the instance matching systems properly. For evaluating scalability, more data can easily be piped from the GeoLink endpoint for additional assessment.

Dataset Preperation We tried to bring freedom and flexibility for system evaluation by combining the datasets and ontologies that we have. In this benchmark, there are four different scenarios for selecting two datasets for an evaluation task. We mapped the R2R and BCO-DMO data to both GMO and GBO ontologies which we discussed in Section 3. As a result, we have four datasets in our benchmark:

- **BCO-DMO_GBO** (BCO-DMO data mapped to GBO ontology)
- **BCO-DMO_GMO** (BCO-DMO data mapped to GMO ontology)
- **R2R_GBO** (R2R data mapped to GBO ontology)
- **R2R_GMO** (R2R data mapped to GMO ontology)

Table 1: Base statistics for the GeoLink Cruise Benchmark

Ontology	Classes	Object Properties	Data Properties	Individuals	Triples
BCO-DMO_GBO	40	149	49	1061	13055
R2R_GBO	40	149	49	5320	27992
BCO-DMO_GMO	79	79	37	1052	16303
R2R_GMO	79	79	37	2025	24798

As a result, in the evaluation of coreference resolution systems, any combination of two datasets from these four datasets could provide a different insight into the IM system and measure different capabilities of them. For instance, two datasets with different ontology schema (R2R_GBO and BCO-DMO_GMO) as the source and target data, could be selected for evaluation tasks or two different datasets with the same ontology schema (BCO-DMO_GMO and R2R_GMO) could be picked up.

Table 1 shows the number of classes and properties in all data sources.

Alignment Tasks Our ground truth data includes manually curated equivalence mappings among 491 entities pairs in BCO-DMO and R2R. In coreference resolution tasks, the alignment system will be asked to find these 491 sameAs links. The performance will be evaluated using precision, recall, and f-measure, as the majority of coreference resolution benchmarks also apply these measures to

¹²https://figshare.com/articles/GeoLink_Cruise_Benchmark/12443999

¹³<http://alignapi.gforge.inria.fr/>

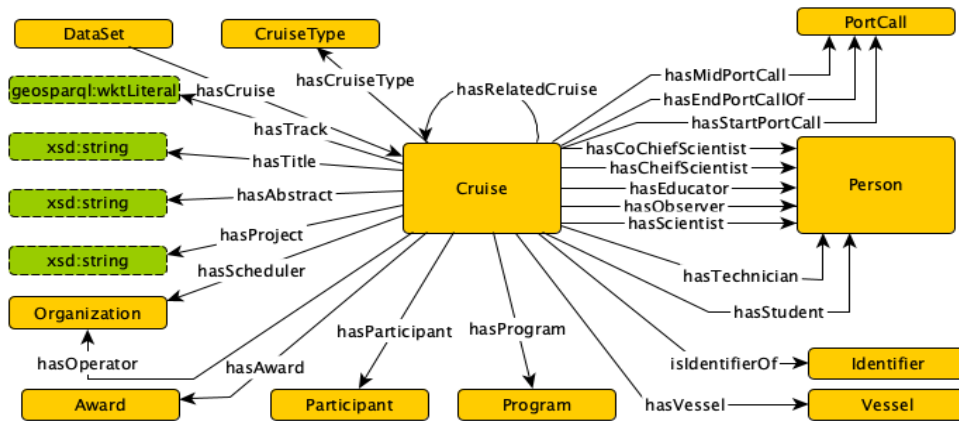


Figure 1: GBO Cruise Ontology Schema

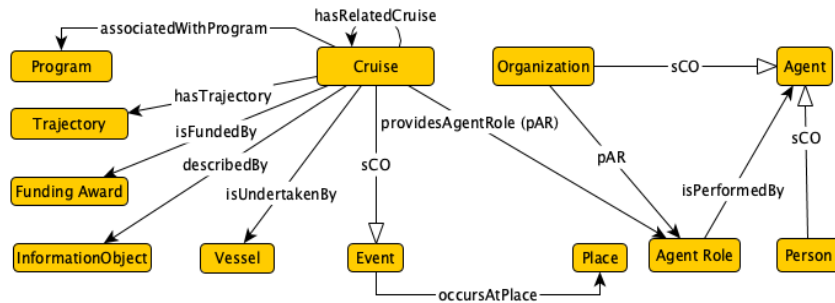


Figure 2: GMO Cruise Ontology Schema, (sCO=subClassOf and pAR=providesAgentRole)

Table 2: owl:sameAs Link Example between Two Cruise Instances in R2R and BCO-DMO.

```

@Prefix bco-dmo: <http://lod.bco-dmo.org/geolink/id/deployment/> .
@Prefix gbo: <https://gbo#> .
@Prefix r2r: <http://data.rvdata.us/id/cruise/> .
@Prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@Prefix foaf: <http://xmlns.com/foaf/0.1/> .
@Prefix rvdata: <http://data.rvdata.us/id/program/> .

```

bco-dmo:57490 bco-dmo:2012	gbo:hasProgram gbo:hasAbstract gbo:hasTitle gbo:hasAcronym	bco-dmo:2012 . "U.S. GLOBEC (GLOBal ocean ECosystems dynamics)..."; "U.S. GLOBal ocean ECosystems dynamics"@en-us ; "U.S. GLOBEC" .
r2r:AT7-21 rvdata:GLOBEC	gbo:hasProgram rdfs:label gbo:description foaf:name	rvdata:GLOBEC . "Global Ocean Ecosystem Dynamics" ; "Global Ocean Ecosystem Dynamics" ; "Global Ocean Ecosystem Dynamics" .

assess the completeness, soundness, and overall matching quality

of the instance matching systems. The alignment tasks between the ontologies are listed as follows:

Evaluation between BCO-DMO_GMO and R2R_GMO
 Evaluation between BCO-DMO_GBO and R2R_GMO
 Evaluation between BCO-DMO_GMO and R2R_GBO
 Evaluation between BCO-DMO_GBO and R2R_GBO

We expect that coreference resolution systems should perform better in finding the links between the entities of the same schema and semantic-aware comparison should be more straightforward. For example, if we want to evaluate the performance of the IM system in founding the similarities between the datasets with the same ontologies, R2R_GMO and BCO-DMO_GMO could be selected because, in this pair, R2R and BCO-DMO both have GMO ontology. In another combination of the data, comparing R2R_GBO and R2R_GMO is also a good scenario for testing the systems regarding their ability to map identical individuals with different schema.

By providing different datasets with different ontologies, we tried to introduce a benchmark capable of evaluating a system properly in different semantic scenarios.

Ontology Schema: GeoLink has two finely developed schemas designed by domain experts, data providers, and ontologists which provide good support for structure-based approaches in coreference resolution algorithms.

In Figures 1 and 2, it shows all the classes and properties related to Cruise in GBO and GMO respectively. Cruise has such an enriched ontology with various properties (object and data property). For simple string similarities, the property value of some data properties, such as `hasTitle`, `hasProject`, and `hasAbstract` in GBO are easily comparable and for fine-grained comparison, more information could be pulled out by traversing through more object properties and data properties in the graph.

In addition, the GeoLink benchmark includes various OWL and OWL2 axioms that are helpful for evaluating the systems that work based on ontologies. Such axioms utilize OWL vocabulary such as

```
owl:sameAs, owl:priorVersion, owl:versionInfo,
owl:imports, owl:unionOf, owl:inverseOf,
owl:complementOf.
```

Beyond label comparison: Although current coreference resolution benchmarks with synthetic data provide only value-based test cases through adding typographical errors to data, the GeoLink Cruise benchmark can capture matches that go beyond similar class label and string matches in data properties due to its real nature and size. Table 2 is an example of two cruises in BCO-DMO respectively R2R that are in fact the same cruise. As it is shown, these two cruises have both a property called `hasProgram` and if we further look in the data for data about these programs, there are properties such as `hasTitle`, `hasAcronym`, `description`, `name`, and `label`, which are pointing to values which are proper strings, making it possible to utilize string similarity metrics. Only systems that are comprehensive enough to look more deeply into the data can capture these kinds of alignments, because the two Cruise instances may not share the same label, but their object properties may share more information. So having real data with challenges beyond data property and label comparison can help to evaluate coreference systems regarding their suitability for realistic data.

5 EXPERIMENTS AND ANALYSIS

In this section, based on the results of the instance matching track at OAEI 2019,¹⁴ we apply AgreementMakerLight (AML) [7] and test it on the proposed benchmark. The AML source code is published on Github,¹⁵ while other systems such as Lily [21] and FTRL-IM [20] are not open source. AML has in fact been actively participating in all the different tracks in OAEI and has achieved excellent performance overall. Therefore, AML can be considered a representative alignment system for the state of the art in the area of instance matching.

As presented in the OAEI result of 2019, AML achieved an f-measure of 86% on SPIMBENCH for both sandbox and mainbox. It performs a proper job of linking objects at both the ontology and instances level. An example of an equivalence link is depicted in Table 3, which AML found between two entities *bbct17:id* and *bbct1261410333:id* in SPIMBENCH. A closer look reveals that these two instances almost share the same triples with some minor changes in one or two properties and property values (changes are in colored text). As we discussed before, the way that SPIMBENCH generates test cases is by adding spelling changes in source data. As you can see here, the string transformation just added some minor changes and most of the data is untouched. For example here the dataProperty `“cw:title”` and `“cw:shortTitle”` in the source data is being changed to `“cw:title0”` and `“cw:title1”` in the target data in SPIMBENCH. As a result, AML or other systems can easily align the two entities since they don’t differ that much from each other considering both property and value. The importance of having real data for evaluation becomes rather obvious from this example and by comparing it with the realistic GeoLink data as displayed, e.g., in Figure 2.

By running AML on our benchmark, the results are different. AML could not generate good coreferences between the instances in the GeoLink benchmark. However, it found some alignments at the ontology level. We show some of the correct and incorrect alignments that AML could find, while it was running on **BCO-DMO_GBO** and **R2R_GMO**.

Example of correct links:

```
<entity1 rdf:resource="http://gbo#hasFamilyName"/>
<entity2 rdf:resource="http://gmo#familyOrSurname"/>
```

```
<entity1 rdf:resource="http://gbo#hasAffiliation"/>
<entity2 rdf:resource="http://gmo#hasPosition"/>
```

```
<entity1 rdf:resource="http://gbo#isPortOf"/>
<entity2 rdf:resource="http://gmo#r_Port"/>
```

```
<entity1 rdf:resource="http://gbo#hasGivenName"/>
<entity2 rdf:resource="http://gmo#firstOrGivenName"/>
```

Example of incorrect links:

```
<entity1 rdf:resource="http://gbo#hasContributor"/>
<entity2 rdf:resource="http://gmo#hasAttribute"/>
```

```
<entity1 rdf:resource="http://gbo#hasCurator"/>
```

¹⁴<https://project-hobbit.eu/challenges/om2019/>

¹⁵<https://github.com/AgreementMakerLight/AML-Project/releases/tag/v3.1>

Table 3: Example of coreference alignment found between two entities of SPIMBENCH by AML algorithm. (colored parts are the transformations that SPIMBENCH made.)

bbct17:id		bbct1261410333:id	
@prefix bbct17: http://www.bbc.co.uk/things/17/		@prefix bbct1261410333: <http://www.bbc.co.uk/ things/1261410333/	
@prefix cw: http://www.bbc.co.uk/ontologies/creativework/		@prefix cw: http://www.bbc.co.uk/ontologies/ creativework/	
@prefix bbct1509989698: <http://www.bbc.co.uk/ things/1509989698/		@prefix bbct1509989698: <http://www.bbc.co.uk/things/ 1509989698/	
rdf:type	cw:Programme	rdf:type	cw:Programme
cw:title	Khalid Mahmood little very provider king until not it accept be or."	cw:title0	Khalid Mahmood little blue provide
cw:shortTitle	millions method competitive we wrestling head sometimes participation turn result	cw:title1	r king until not it accept be or.
cw:category	bbcc:PoliticsPersons-Additional	cw:category	http://www.bbc.co.uk/category/ Politic-sPersonsAdditional
cw:description	" craftsmen make maintained pay layout allowing as commerce did politics see." .	cw:description	" craftsmen make maintained pay layout allowing as commerce did politics see."
cw:about	wikidata:Q695028	cw:about	wikidata:Q695028
cw:about	http://dbpedia.org/resource /Cypriot-PresidentialElection2003	cw:about	http://dbpedia.org/resource/ Cypriot-presidentialelection.2003
cw:mentions	http://sws.geonames.org /7299636	cw:mentions	http://sws.geonames.org /7299636/
cw:audience	cw:InternationalAudience	cw:audience	cw:InternationalAudience
cw:liveCoverage	true http://www.w3.org /2001/XMLSchema/boolean	cw:liveCoverage	true http://www.w3.org/2001/ XMLSchema/boolean
cw:primary Format	- cw:AudioFormat	cw:primary Format	- cw:AudioFormat
cw:dateCreated	"2011-11-17T20:19:06.076+02:00" http://www.w3.org/2001/ XMLSchema/dateTime	cw:dateCreated	"2011-11-17T20:19:06.076+02:00" http://www.w3.org/2001/ XMLSchema/dateTime
cw:dateModified	"2012-10-10T04:09:55.770+03:00" http://www.w3.org/2001/ XMLSchema/dateTime	cw:dateModified	"2012-10-10T04:09:55.770+03:00" http://www.w3.org/2001/ XMLSchema/dateTime
cw:thumbnail	bbct17:219976210	cw:thumbnail	bbct:219976210
cw:altText	"thumbnail atlText for CW http://www.bbc.co.uk/ context/17/id		
bbc:primary-ContentOf	bbct1509989698:id	bbc:primary-ContentOf	bbct1509989698:id

<entity2 rdf:resource="http://www.w3.org/2006/time #hasDuration"/>

As we have seen, AML found property alignments between GMO and GBO. The system is working fine on finding the links that have a higher string similarity or string synonyms between the two objects, but for the objects with weaker string similarities like "hasAttribute" and "hasContribute", AML usually reported false positives by wrongly adding a mapping between the objects with partial string matching.

In conclusion, having a simple synthetic benchmark might be useful for the evaluation of the systems in specific areas in an early stage. A benchmark comprising real-world data can boost the development of the state of the art alignment algorithms in the field of instance matching.

6 CONCLUSION AND FUTURE WORKS

Coreference resolution on semantically linked data has been discussed for over a decade now, but not much work has been done related to a realistic benchmark for systematic evaluation of these systems. In this paper, we have proposed the use of the GeoLink Cruise data repository as a coreference resolution benchmark. The GeoLink Cruise benchmark utilizes manually curated equivalents

links between Cruise instances of R2R and BCO-DMO data. This Benchmark has been accepted by OAEI 2020 and is going to exist for evaluating systems going forward.

An analysis of current automated coreference resolution systems indicates that current synthetic benchmarks are insufficiently replicating the complexities present in real-world data, and our dataset can be used to evaluate entity resolution approaches from a more realistic perspective. Additionally, our dataset has the potential to spur innovations related to scalability and data aligning with geospatial and temporal aspects.

In the future, we plan to actively keep maintaining the benchmark and contribute to its use at the OAEI. In addition, we also plan to assess the performance of several existing approaches to coreference resolution on this benchmark. Furthermore, we intend to develop an alignment system that can meet the challenges proposed by our benchmark.

ACKNOWLEDGMENTS

Lu Zhou and Pascal Hitzler acknowledge support by the National Science Foundation under Grant 1936677 through *Convergence Accelerator Phase I (RAISE): Spatially-Explicit Models, Methods, and Services for Open Knowledge Networks*.

REFERENCES

- [1] Samur Araújo, Arjen P. de Vries, and Daniel Schwabe. 2011. SERIMI results for OAEI 2011. In *Proceedings of the 6th International Workshop on Ontology Matching, Bonn, Germany, October 24, 2011 (CEUR Workshop Proceedings, Vol. 814)*, Pavel Shvaiko, Jérôme Euzenat, Tom Heath, Christoph Quix, Ming Mao, and Isabel F. Cruz (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-814/oeai11_paper15.pdf
- [2] Michelle Cheatham, Reihaneh Amini, and Chandan Patel. 2016. Matching instances in GeoLink. In *Proceedings of the 11th International Workshop on Ontology Matching co-located with the 15th International Semantic Web Conference (ISWC 2016), Kobe, Japan, October 18, 2016 (CEUR Workshop Proceedings, Vol. 1766)*, Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Michelle Cheatham, Oktie Hassanzadeh, and Ryutaro Ichise (Eds.). CEUR-WS.org, 237–238. http://ceur-ws.org/Vol-1766/om2016_poster6.pdf
- [3] Michelle Cheatham, Adila Krisnadhi, Reihaneh Amini, Pascal Hitzler, Krzysztof Janowicz, Adam Shepherd, Tom Narock, Matt Jones, and Peng Ji. 2018. The GeoLink knowledge graph. *Big Earth Data* 2, 2 (2018), 131–143.
- [4] Isabel F. Cruz, Cosmin Stroe, Federico Caimi, Alessio Fabiani, Catia Pesquita, Francisco M. Couto, and Matteo Palmonari. 2011. Using AgreementMaker to align ontologies for OAEI 2011. In *Proceedings of the 6th International Workshop on Ontology Matching, Bonn, Germany, October 24, 2011 (CEUR Workshop Proceedings, Vol. 814)*, Pavel Shvaiko, Jérôme Euzenat, Tom Heath, Christoph Quix, Ming Mao, and Isabel F. Cruz (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-814/oeai11_paper1.pdf
- [5] Syrine Damak, Hazem Souid, Marouen Kachroudi, and Sami Zghal. 2015. EX-ONA results for OAEI 2015. In *Proceedings of the 10th International Workshop on Ontology Matching collocated with the 14th International Semantic Web Conference (ISWC 2015), Bethlehem, PA, USA, October 12, 2015 (CEUR Workshop Proceedings, Vol. 1545)*, Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Michelle Cheatham, and Oktie Hassanzadeh (Eds.). CEUR-WS.org, 145–149. http://ceur-ws.org/Vol-1545/oeai15_paper5.pdf
- [6] Jérôme David, Jérôme Euzenat, François Scharffe, and Cássia Trojahn dos Santos. 2011. The Alignment API 4.0. *Semantic Web* 2, 1 (2011), 3–10. <https://doi.org/10.3233/SW-2011-0028>
- [7] Daniel Faria, Catia Pesquita, Emanuel Santos, Isabel F. Cruz, and Francisco M. Couto. 2013. AgreementMakerLight results for OAEI 2013. In *Proceedings of the 8th International Workshop on Ontology Matching co-located with the 12th International Semantic Web Conference (ISWC 2013), Sydney, Australia, October 21, 2013 (CEUR Workshop Proceedings, Vol. 1111)*, Pavel Shvaiko, Jérôme Euzenat, Kavitha Srinivas, Ming Mao, and Ernesto Jiménez-Ruiz (Eds.). CEUR-WS.org, 101–108. http://ceur-ws.org/Vol-1111/oeai13_paper1.pdf
- [8] Alfio Ferrara, Andriy Nikolov, Jan Noessner, and François Scharffe. 2013. Evaluation of instance matching tools: The experience of OAEI. *J. Web Semant.* 21 (2013), 49–60. <https://doi.org/10.1016/j.websem.2013.05.004>
- [9] Md. Seddiqui Hanif and Masaki Aono. 2009. Anchor-Flood: 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 (CEUR Workshop Proceedings, Vol. 551)*, Pavel Shvaiko, Jérôme Euzenat, Fausto Giunchiglia, Heiner Stuckenschmidt, Natalya Fridman Noy, and Arnon Rosenthal (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-551/oeai09_paper1.pdf
- [10] Pascal Hitzler, Aldo Gangemi, Krzysztof Janowicz, Adila Krisnadhi, and Valentina Presutti (Eds.). 2016. *Ontology Engineering with Ontology Design Patterns - Foundations and Applications*. Studies on the Semantic Web, Vol. 25. IOS Press.
- [11] Wei Hu, Jianfeng Chen, Gong Cheng, and Yuzhong Qu. 2010. ObjectCoref & Falcon-AO: results for OAEI 2010. In *Proceedings of the 5th International Workshop on Ontology Matching (OM-2010), Shanghai, China, November 7, 2010 (CEUR Workshop Proceedings, Vol. 689)*, Pavel Shvaiko, Jérôme Euzenat, Fausto Giunchiglia, Heiner Stuckenschmidt, Ming Mao, and Isabel F. Cruz (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-689/oeai10_paper6.pdf
- [12] Jakob Huber, Timo Szttyler, Jan Nöbner, and Christian Meilicke. 2011. CODI: Combinatorial Optimization for Data Integration: results for OAEI 2011. In *Proceedings of the 6th International Workshop on Ontology Matching, Bonn, Germany, October 24, 2011 (CEUR Workshop Proceedings, Vol. 814)*, Pavel Shvaiko, Jérôme Euzenat, Tom Heath, Christoph Quix, Ming Mao, and Isabel F. Cruz (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-814/oeai11_paper4.pdf
- [13] Abderrahmane Khia, Moussa Benaissa, and Mohammed Amine Belfedhal. 2015. STRIM results for OAEI 2015 instance matching evaluation. In *Proceedings of the 10th International Workshop on Ontology Matching collocated with the 14th International Semantic Web Conference (ISWC 2015), Bethlehem, PA, USA, October 12, 2015 (CEUR Workshop Proceedings, Vol. 1545)*, Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Michelle Cheatham, and Oktie Hassanzadeh (Eds.). CEUR-WS.org, 208–215. http://ceur-ws.org/Vol-1545/oeai15_paper16.pdf
- [14] Adila Alfa Krisnadhi, Yingjie Hu, Krzysztof Janowicz, Pascal Hitzler, Robert A. Arko, Suzanne Carbotte, Cynthia Chandler, Michelle Cheatham, Douglas Fils, Tim Finin, Peng Ji, Matthew B. Jones, Nazifa Karima, Kerstin A. Lehnert, Audrey Mickle, Tom Narock, Margaret O'Brien, Lisa Raymond, Adam Shepherd, Mark Schildhauer, and Peter Wiebe. 2015. 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 (CEUR Workshop Proceedings, Vol. 1486)*, Serena Villata, Jeff Z. Pan, and Mauro Dragoni (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-1486/paper_99.pdf
- [15] Khai Nguyen and Ryutaro Ichise. 2013. SLINT+ results for OAEI 2013 instance matching. In *Proceedings of the 8th International Workshop on Ontology Matching co-located with the 12th International Semantic Web Conference (ISWC 2013), Sydney, Australia, October 21, 2013 (CEUR Workshop Proceedings, Vol. 1111)*, Pavel Shvaiko, Jérôme Euzenat, Kavitha Srinivas, Ming Mao, and Ernesto Jiménez-Ruiz (Eds.). CEUR-WS.org, 177–183. http://ceur-ws.org/Vol-1111/oeai13_paper11.pdf
- [16] Xing Niu, Shu Rong, Yunlong Zhang, and Haofen Wang. 2011. Zhishi.links results for OAEI 2011. In *Proceedings of the 6th International Workshop on Ontology Matching, Bonn, Germany, October 24, 2011 (CEUR Workshop Proceedings, Vol. 814)*, Pavel Shvaiko, Jérôme Euzenat, Tom Heath, Christoph Quix, Ming Mao, and Isabel F. Cruz (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-814/oeai11_paper16.pdf
- [17] Fatiha Sais, Nibal B. Niraula, Nathalie Pernelle, and Marie-Christine Rousset. 2010. LN2R: a knowledge based reference reconciliation system: OAEI 2010 results. In *Proceedings of the 5th International Workshop on Ontology Matching (OM-2010), Shanghai, China, November 7, 2010 (CEUR Workshop Proceedings, Vol. 689)*, Pavel Shvaiko, Jérôme Euzenat, Fausto Giunchiglia, Heiner Stuckenschmidt, Ming Mao, and Isabel F. Cruz (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-689/oeai10_paper8.pdf
- [18] Tzanina Saveta, Evangelia Daskalaki, Giorgos Flouris, Irini Fundulaki, Melanie Herschel, and Axel-Cyrille Ngonga Ngomo. 2015. Pushing the Limits of Instance Matching Systems: A Semantics-Aware Benchmark for Linked Data. In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18–22, 2015 - Companion Volume*, Aldo Gangemi, Stefano Leonardi, and Alessandro Panconesi (Eds.). ACM, 105–106. <https://doi.org/10.1145/2740908.2742729>
- [19] Aynaz Taheri and Mehrnosh Shamsfard. 2012. SBUEL: results for OAEI 2012. In *Proceedings of the 7th International Workshop on Ontology Matching, Boston, MA, USA, November 11, 2012 (CEUR Workshop Proceedings, Vol. 946)*, Pavel Shvaiko, Jérôme Euzenat, Anastasios Kementsietsidis, Ming Mao, Natasha Fridman Noy, and Heiner Stuckenschmidt (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-946/oeai12_paper11.pdf
- [20] Xiaowen Wang, Yizhi Jiang, Yi Luo, Hongfei Fan, Hua Jiang, Hongming Zhu, and Qin Liu. 2019. FTRLIM Results for OAEI 2019. In *Proceedings of the 14th International Workshop on Ontology Matching co-located with the 18th International Semantic Web Conference (ISWC 2019), Auckland, New Zealand, October 26, 2019 (CEUR Workshop Proceedings, Vol. 2536)*, Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn (Eds.). CEUR-WS.org, 146–152. http://ceur-ws.org/Vol-2536/oeai19_paper9.pdf
- [21] Jiangheng Wu, Zhe Pan, Ce Zhang, and Peng Wang. 2019. Lily Results for OAEI 2019. In *Proceedings of the 14th International Workshop on Ontology Matching co-located with the 18th International Semantic Web Conference (ISWC 2019), Auckland, New Zealand, October 26, 2019 (CEUR Workshop Proceedings, Vol. 2536)*, Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn (Eds.). CEUR-WS.org, 153–159. http://ceur-ws.org/Vol-2536/oeai19_paper10.pdf
- [22] Katrin Zaiß, Stefan Conrad, and Sven Vater. 2010. A Benchmark for Testing Instance-based Ontology Matching Methods. In *Proceedings of the EKAW2010 Poster and Demo Track, Lisbon, Portugal, October 11 - 15, 2010 (CEUR Workshop Proceedings, Vol. 674)*, Johanna Völker and Óscar Corcho (Eds.). CEUR-WS.org. <http://ceur-ws.org/Vol-674/Paper157.pdf>
- [23] Yan Zhang, Hailong Jin, Liangming Pan, and Juan-Zi Li. 2016. RiMOM results for OAEI 2016. In *Proceedings of the 11th International Workshop on Ontology Matching co-located with the 15th International Semantic Web Conference (ISWC 2016), Kobe, Japan, October 18, 2016 (CEUR Workshop Proceedings, Vol. 1766)*, Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Michelle Cheatham, Oktie Hassanzadeh, and Ryutaro Ichise (Eds.). CEUR-WS.org, 210–216. http://ceur-ws.org/Vol-1766/oeai16_paper13.pdf
- [24] Lu Zhou, Michelle Cheatham, Adila Krisnadhi, and Pascal Hitzler. 2018. A Complex Alignment Benchmark: GeoLink Dataset. In *The Semantic Web - ISWC 2018 - 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 11137)*, Denny Vrandečić, Kalina Bontcheva, Mari Carmen Suárez-Figueroa, Valentina Presutti, Irene Celino, Marta Sabou, Lucie-Aimée Kaffee, and Elena Simperl (Eds.). Springer, 273–288. https://doi.org/10.1007/978-3-030-00668-6_17
- [25] Lu Zhou, Michelle Cheatham, Adila Krisnadhi, and Pascal Hitzler. 2020. GeoLink Data Set: A Complex Alignment Benchmark from Real-world Ontology. *Data Intell.* 2, 3 (2020), 353–378. https://doi.org/10.1162/dint_a_00054

Towards Complex Ontology Alignment using Large Language Models

Reihaneh Amini¹, Sanaz Saki Norouzi¹, Pascal Hitzler¹, and Reza Amini²

¹ Kansas State University, Manhattan, KS, USA
{reihanea,sanazsn,hitzler}@ksu.edu

² Wright State University, Dayton, OH, USA amini.4@wright.edu

Abstract. Ontology alignment, a critical process in the Semantic Web for detecting relationships between different ontologies, has traditionally focused on identifying so-called “simple” 1-to-1 relationships through class labels and properties comparison. The more practically useful exploration of more complex alignments remains a hard problem to automate, and as such is largely underexplored, i.e. in application practice it is usually done manually by ontology and domain experts. Recently, the surge in Natural Language Processing (NLP) capabilities, driven by advancements in Large Language Models (LLMs), presents new opportunities for enhancing ontology engineering practices, including ontology alignment tasks. This paper investigates the application of LLM technologies to tackle the complex ontology alignment challenge. Leveraging a prompt-based approach and integrating rich ontology content – so-called *modules* – our work constitutes a significant advance towards automating the complex alignment task.

Keywords: Complex Ontology Alignment · Ontology · Large Language Model · Knowledge Graph · Modular Ontology Modeling

1 Introduction

Ontology alignment (sometimes called ontology matching) [8] is the task of establishing mappings between different ontologies, and as a research field it is concerned with ways to automate or at least semi-automate this task. For those not familiar with the field: Ontologies, which are usually knowledge bases expressed using Description Logics [12] (including the W3C Web Ontology Language – OWL – standard [22]) in this case act as a type of data schema for data expressed as knowledge graphs [11], i.e. the establishing of mappings between ontologies is central for schema-based data integration purposes.

Ontology alignment has been studied for over two decades, resulting in the development of many alignment approaches and systems. The majority of these systems are designed to detect only so-called “simple” 1-to-1 mappings between ontologies, primarily by establishing equivalence relationships between classes (unary predicates), or between properties (binary relationships); for example, one ontology may have a class called “Person” while another may have a class

called “Human”, and an ontology alignment mapping may state that these two classes are in fact equivalent. It has long been recognized in the Semantic Web and Ontologies community that such simple mappings are helpful but ultimately insufficient for data integration task, for which mappings would need to be in the form of complex mapping rules³, that can be expressed, e.g., as Datalog rules. However, detecting complex alignments between ontologies remains a very challenging and thus largely unexplored area with only few contributions that made progress in restricted settings (see Section 2). In current practice, establishing complex alignments between two or more ontologies requires domain experts to collaborate and manually generate the alignments, and this is usually a very work-intensive and thus expensive task. Any automation or semi-automation would have significant added value.

The Ontology Alignment Evaluation Initiative (OAEI)⁴ is a long-standing coordinated international effort aimed at improving and evaluating ontology alignment and coreference resolution technologies.⁵ It organizes annual evaluation campaigns [20] that provide a controlled environment where participants can test their ontology alignment systems using various benchmark tests. The benchmarks cover a range of complexity levels and real-world scenarios, aiming to simulate different aspects of the ontology alignment process.

With significant advancements in the natural language processing (NLP) and natural language understanding (NLU) fields, spurred by Large Language Models (LLMs), it has become possible to extract meanings from text and reason about it more effectively. OpenAI⁶ has been at the forefront of this research, developing the Generative Pre-trained Transformer (GPT) series of models, which have attracted considerable attention from researchers, developers, and users. One of the most notable models, introduced in March 2022, was GPT-4 [1]. This transformer-based model is designed to predict the next token in generating text and has shown improvements in producing results that more closely align with user intent compared to its predecessor, GPT-3.5, on 70.2% of the prompts.⁷

Recent advancements in applying Large Language Models (LLMs) to Semantic Web and ontology engineering tasks have shown promising results, due to the importance of NLP for such tasks. A notable study [10] demonstrated the effectiveness of zero-shot and few-shot prompting with LLMs on various tasks within the Ontology Alignment Evaluation Initiative (OAEI), highlighting their potential in this area. The study was restricted to simple alignments. Indeed, as we will see later, a straightforward tasking of LLMs with the production of complex alignments does not quite work.

³ See Section 4 for an example.

⁴ <http://oaei.ontologymatching.org/>

⁵ Co-reference resolution is about establishing equivalence and non-equivalence mappings between individuals, or constants, a related but different task of similar practical importance.

⁶ <https://openai.com>

⁷ See OpenAI: Introducing ChatGPT, 2022, <https://openai.com/blog/chatgpt> and Greg Brockman, Peter Welinder, Mira Murati, and OpenAI, 2020, <https://openai.com/blog/openai-api>.

One of the difficulties with ontology alignment is that ontologies often tend to be underspecified and with little internal structure that may add some self-explainability. This can be seen for example by the considerable disagreement between humans as to “correct” alignments, even for the simple alignment task [6]. It has been posited that additional internal structure, e.g. in the form of conceptual “ontology modules” should aid with ontology engineering tasks that are hard to automate [23]. Following this hitherto merely conceptual argument, we made use of ontology modules in our approach to generate complex alignments, and as we will report below, our prompt-based approach for discovering complex alignments between ontologies yields significantly better results when richer content in the form of ontology modules is available.

Since ontologies are knowledge bases expressed using formal logic, and mapping rules are also expressed using formal logic and processed as such, ontology alignment is a key symbolic task that we are here addressing using “neural” means (i.e. LLMs as artificial neural networks), as such contributing to the body of approaches and methods for neural-symbolic learning and reasoning [13].

The structure of this paper is as follows: Section 2 reviews past research and the current state-of-the-art methods in complex alignment. Section 3 outlines the methodologies and approaches we utilized to experiment with an effective prompt-based method for discovering complex alignments. The evaluation of our approach and the specifics of prompt tailoring are discussed in Sections 4 and 5, respectively.

2 Related Work

In the semantic web domain, ontology alignment is crucial to discerning correspondences between concepts and properties across various ontologies. It involves two primary types of matching: simple and complex alignment [24]. Simple alignment focuses on identifying 1-to-1 equivalence relationships between concepts in two different ontologies. Complex alignment aims to match 1-to-n, n-to-1, or m-to-n equivalence or subsumption relationships between complex expressions in two ontologies (see e.g. [28]); the alignments are usually expressed in form of (first-order logic) rules, description logic axioms, or similar logical or quasi-logical notation.

Most of the work on ontology alignment is evaluated by the OAEI, an annual event focusing on ontology-matching systems’ performance, starting in 2004, with focus mostly on simple alignment. A complex alignment track was started in 2018 because simple matching is often not sufficient to capture the rich semantics needed for advanced applications using realistic ontologies. In the context of this track, the datasets utilized throughout these years encompass Complex Conference, Populated Complex Conference, Hydrography, GeoLink, Populated GeoLink, Populated Enslaved, and Taxon. It should be noted, however, that some datasets have ceased to be evaluated in recent years [2,28,20] as each track depends on volunteers to run it.

The GeoLink complex alignment benchmark stands out as a real-world dataset for testing complex ontology alignment; using real-world datasets like GeoLink for testing tasks such as ontology alignment is crucial as it reflects the complexities of practical applications, helps validate the robustness and versatility of ontology alignment algorithms, and ensures their effectiveness beyond synthetic ones. It contains two ontologies, the GeoLink Base Ontology (GBO) and the GeoLink Modular Ontology (GMO). These are paired with a reference alignment created with help from domain specialists [28,4]. During these years, only two or three systems registered for evaluation in this track. Evaluation results show [3,19,18] that the systems that make use of instance data that is shared between the two ontologies perform relatively well, with high precision but low recall, i.e. many matches are missed. Furthermore, the availability of shared instance data, as for this benchmarking competition, while helpful for advancing the state of the art, is unrealistic for most practical application scenarios for ontology alignment: Usually, such shared data would not be available. The new results which we present in this paper have been obtained without taking shared instance data into consideration, which, in our opinion, constitutes a major advance over the state of the art.

Currently, there is a plethora of research on applications of LLMs in a wide variety of fields, often with significant success. In the field of ontology alignment, [17] presents a zero-shot evaluation of ChatGPT-4 using the OAEI conference track ontologies (consisting of small and medium sized ontologies), showing a high recall and F1 of 0.52. Also, [9] provides a zero-shot evaluation of GPT 3.5 and Flan-T5-XXL on the OAEI Bio-ML track, and they also achieved low precision and high recall. [10] employs a comparable approach implemented in MELT, and the candidate generation and matching processes work in collaboration with Sentence-BERT, their evaluation has been done on some OAEI track datasets which almost achieve high F1 scores. In [21], they present a methodology utilizing LLM-based agents for retrieval and matching processes, which they enhance by incorporating Retrieval Augmented Generation (RAG) within their agents. The assessment was conducted on several tracks of OAEI. The results from the conference track indicate a high recall but low precision. The performance of the proposed method on other tracks is relatively good. However all these results do not go beyond simple alignment. In particular, no good results for complex alignment benchmarks have been reported yet.

3 Complex Alignment by Large Language Model

3.1 Ontology Modules

Ontologies that are conceptually clear are inherently more reusable, primarily because they are straightforward to comprehend from the outset. Consequently, a primary objective in the field of ontology research is to develop ontology modeling methodologies for creating ontologies with high conceptual clarity. It is a

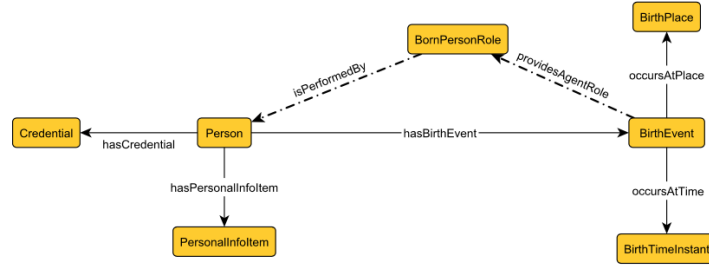


Fig. 1. A diagram for a Person module

fundamental assumption that adhering to well-established modeling principles enhances the understandability and reusability of an ontology [23].

It is known that repurposing or adjusting of ontologies for new applications often presents challenges. These challenges stem from various factors, including: (i) mismatches in the level of detail between ontologies and the intended use-cases, (ii) unclear concepts within the ontologies intended for reuse, (iii) the absence and the complexity of ontology alignment practices, and (iv) the scarcity of tools designed to facilitate ontology reuse and support the process. To tackle these issues, the Modular Ontology Modeling (MOMo) approach was proposed [23]. In the MOMo framework, a key feature is the creation and connection of compact, independent modules, which offers numerous advantages. For instance, it streamlines maintenance because changes to a single module have little to no effect on the entire ontology. Additionally, tracking the origin of every ontology segment back to the initial requirements becomes straightforward through the use of module documentation or metadata.

In the Semantic Web community, the term “**module**” can mean many things. In our case, it refers to a specific part of an ontology that encapsulates a principal concept and its main features, such as a Person module capturing details about “BirthEvent”, “PersonalInfoItem”, and “Credential” as depicted in Figure 1. Modules serve dual roles: they are technical constructs that, on the one hand, demarcate parts of ontologies that group related classes and their interactions and, on the other hand, they do this in a way that aligns with domain experts’ understanding. Despite potential overlaps and hierarchical structures within them, modules organize the ontology into a network of interrelated pieces, each mirroring the domain’s conceptual framework as understood by experts.

Modules enable a strategic approach to ontology modeling by allowing the work to be broken down into manageable segments; initially focusing on individual modules before linking them together. This approach provides a clear way to manage the complexities of large, cohesive ontologies by breaking down the process into understanding individual modules and their interconnections. This modular approach also aligns with the way domain experts conceptualize their fields, making both the ontology and its documentation more accessible and comprehensible. Since each module can be easily swapped out for another—perhaps

one that offers a different level of detail—modifications are localized, making the entire system more adaptable [15].

In the research we present herein, we explore the impact of integrating module information on the effectiveness of identifying complex alignments. Specifically, we include descriptions, Core Axioms (where applicable), and Alignment information as outlined in the GeoLink Core Ontology Design Patterns [14] that constitute the GeoLink Modular Ontology (GMO), that had been developed as an integrated schema for combining several large-scale ocean science data repositories [7]. As we will see further below, the utilization of additional module information is of core importance in order to solve the complex alignment problem, in our setting.

3.2 Design of the Prompting Process

Tailoring an LLM for specific tasks can be achieved either by fine-tuning the model with select data or by using prompts for in-context learning. Fine-tuning requires significant computational resources and expertise, which may not be feasible for all users or organizations. Moreover, it risks over-fitting the model to the training data, potentially diminishing its ability to generalize to other tasks. In contrast, employing various prompting techniques on a fixed LLM is more resource-efficient, requiring less computational power, time, and expertise. Additionally, prompts allow for the flexible and on-the-fly adaptation of the model to a wide range of tasks by applying different prompting techniques [16].

Prompt engineering is increasingly vital for enhancing the performance of large language models (LLMs) across a wide range of tasks. These models have shown impressive capabilities using zero-shot prompts, where they generate responses without prior specific training examples. However, for more complex tasks requiring deeper reasoning, advanced prompting techniques are necessary. Incorporating context into prompts significantly improves model performance. Techniques such as few-shot prompting, where the model is given a few examples to learn from, or chain-of-thought prompting, which guides the model through a series of reasoning steps, are particularly effective in enhancing the model’s understanding and response accuracy [26].

There is a range of task-agnostic prompting techniques available for use with large language models.⁸ These include:

- Zero-shot prompting: This technique involves providing a single, natural language description of the task at inference time, without any prior examples.
- Few-shot prompting: This approach includes giving the model a few examples of the task, complete with context and successful outcomes, to guide its understanding and responses [5].
- Chain-of-thought prompting: This method involves supplying the model with a series of thought process examples, helping it to navigate through reasoning steps to arrive at a conclusion [25].

⁸ <https://www.promptingguide.ai/techniques>

These techniques enable large language models to adapt and respond to a wide variety of tasks with varying levels of guidance and specificity.

Figure 2 illustrates our workflow, which begins by uploading the entire GMO file as an initial prompt. Subsequently, we include specific entities from the second ontology, the GBO, to inquire about their alignment with the GMO. If the initial prompt successfully provides the relevant segments, further prompts in the chain-of-thought process are unnecessary. Otherwise, we can provide a list of all GMO module names and ask GPT to identify the most related modules for the GBO entities. Then, in a subsequent prompt, we supplement the inquiry with module information and request the related segments again.

In the following section, we examine the performance of GPT-4 in response to various entities, modules, and prompts it processes.

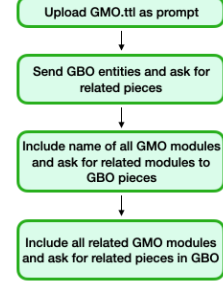


Fig. 2. Prompting workflow

4 Evaluation

For the evaluation of our methodology, we employ the GeoLink Complex Alignment dataset previously detailed in Zhou et al., 2018 [27]. This dataset comprises 109 complex alignment rules between GMO and GBO, each of which has been thoroughly analyzed to test our hypothesis. Detailed findings from this study, including both the prompts used and the responses obtained, are available online.⁹

4.1 Prompting and Detailed Example

OpenAI’s functionality includes the ability to load a prompt from a file, triggering backend processing to parse the uploaded data. We uploaded the entire GMO ontology RDF file in turtle (.ttl) format. OpenAI GPT-4 parsed the data, albeit with some latency, prompting a request for more specific tasks to be directed to it (Figure 3).

The complex alignment rule we are targeting in this example is

$$\begin{aligned}
 &\text{Award}(x) \wedge \text{hasCoPrincipalInvestigator}(x, z) \leftrightarrow \\
 &\text{FundingAward}(x) \wedge \text{providesAgentRole}(x, y) \\
 &\wedge \text{CoPrincipalInvestigatorRole}(y) \wedge \text{performedBy}(y, z)
 \end{aligned} \tag{1}$$

where GBO entities `hasCoPrincipalInvestigator` (a relation or binary predicate or so-called object property) and `Award` (a class or unary predicate)

⁹ <https://daselab.cs.ksu.edu/publications/alignment-rules-gbo-gmo>

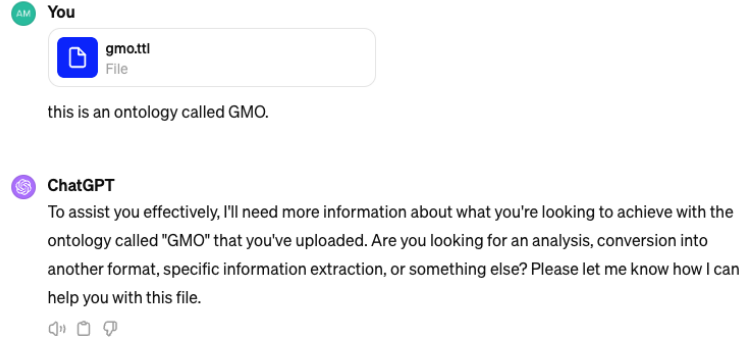


Fig. 3. Uploading GMO.ttl file as prompt.

can be found on the left-hand side, and the GMO entities on the right-hand side. This is an example of a complex alignment rule expressed in first-order predicate logic and our objective is to assess the effectiveness of detecting complex alignments by prompting GPT-4.

After uploading the GMO, we extract the GBO entities found on the left-hand side of the complex alignment rule (i.e., “**Award**” and “**hasCoPrincipalInvestigator**”) from the GBO RDF file in its original format. For our example, this looks as follows.

```
### http://gbo#Award
main:Award rdf:type owl:Class ;
          rdfs:comment "Funding provided by an Organization
enabling Participation.";
          rdfs:label "Award" .

### http://gbo#hasCoPrincipalInvestigator
main:hasCoPrincipalInvestigator rdf:type owl:ObjectProperty ;
                                owl:inverseOf main:isCoPrincipalInvestigatorOf ;
                                rdfs:domain [ rdf:type owl:Class ;
                                              owl:unionOf ( main:Award
                                                            main:Program
                                                            )
                                              ] ;
                                rdfs:range main:Person ;
                                rdfs:label "hasCoPrincipalInvestigator" .
```

Next, our prompt instructs GPT-4 to examine the components in GMO that are associated with these elements in GBO, as depicted in Figure 4.

What we observed from GPT-4’s responses in this step is often not a comprehensive answer. Many times, it explicitly states that *It appears there were no results found in the ontology file “gmo.ttl” directly related to the terms you mentioned from the “GBO” ontology, such as “Award”, “hasCoPrincipalInvestigator”, “Person”, “Program”, or “isCoPrincipalInvestigatorOf”*. GPT-4 typically continues its response with further inquiries such as *To proceed, I can*

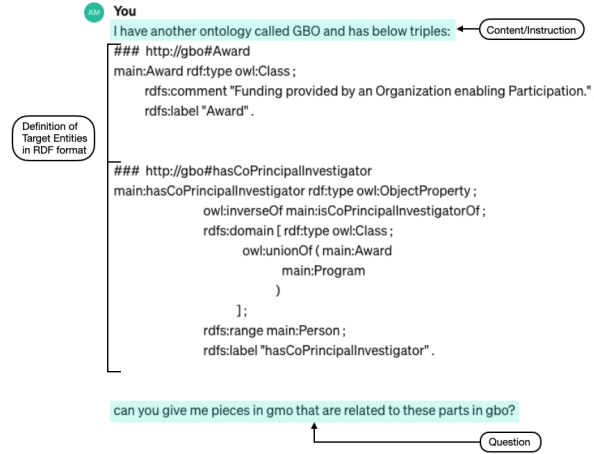


Fig. 4. GBO-related instruction and question in prompt

manually examine the “gmo.ttl” file to identify any related concepts or properties that might align with those in the “GBO” ontology. This process involves looking for classes or properties that serve similar purposes or are conceptually related, even if they are not directly named the same. Would you like me to perform this manual examination, or is there another way I can assist you with these ontologies?

Our consistent response to this question has always been “Yes”, and the typical output we received from GPT is illustrated in Fig 5. As depicted, the LLM successfully identifies a similar entity to **gbo#Award**, which in GMO is **gmo#AwardAmount** (highlighted in green). However, it struggles to conduct further analysis to identify alignments in GMO related to the relationship between “Award” and “hasCoPrincipalInvestigator” in GBO. Rather, it attempts to generate random information about recently detected entities such as “FundingAward”, which hypothetically arises from its next-word-prediction logic (highlighted in yellow). However, the remaining generated information, such as “gmo#Program” (highlighted in red), is not relevant to this component in GBO. Ultimately, it states that *There were no direct matches found for hasCoPrincipalInvestigator or isCoPrincipalInvestigatorOf in the GMO ontology.*

We also attempted zero-shot prompting at this stage, combining both the GMO file and the second prompt into a single input to GPT-4. This approach resulted in increased latency and a confused response.

Our conclusion from this analysis is that GPT-4 failed to detect the complex alignment between the two ontologies. Instead, it provided partially related objects, including random information about different unrelated entity classes.

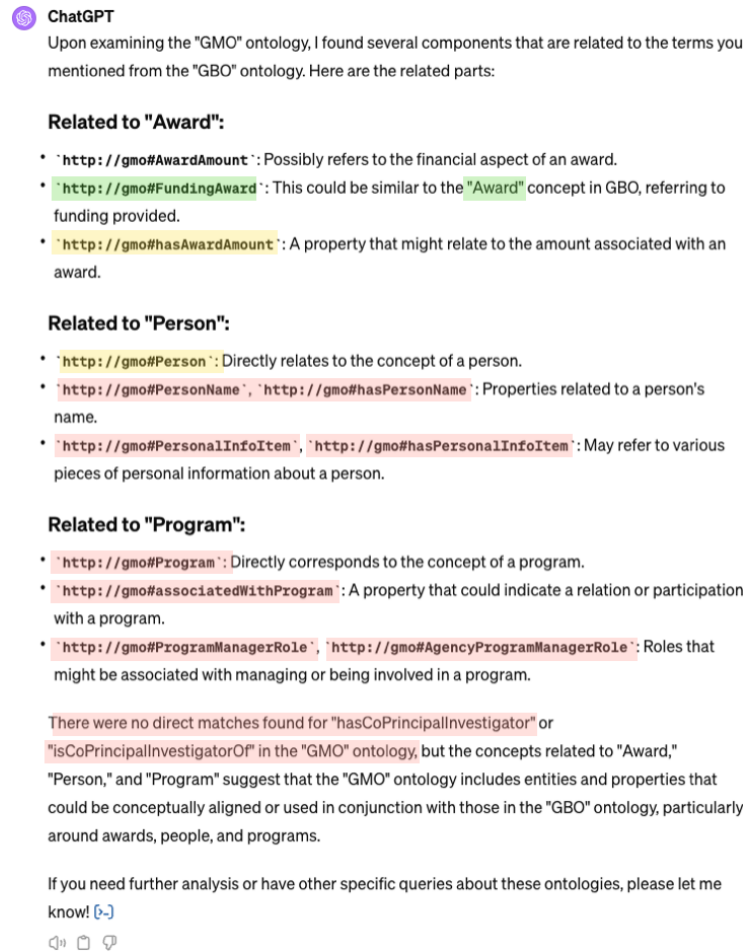


Fig. 5. GPT-4 response to our initial question

This is a list of all the modules and patterns that are defined in GMO ontology:

Agent , Agent Role, Event, Information Object, Identifier, Person, Personal Info Item pattern, Person Name, Organization, Funding Award, Program, Place, Cruise, Platform Pattern Stub, Vessel pattern, Physical Sample pattern, Property Value pattern.

From this list, give me relevant module names that can help us find the most related GMO pieces.

Fig. 6. Prompt for module name suggestion

To proceed, let us revisit the GMO modules discussed in Section 3.1. The documentation [14] contains an informal description of the modules (in the documentation called *patterns*) in the GMO ontology, accompanied by visual depictions of these patterns. In the ongoing chain-of-thoughts prompting processes, we included the module names and asked for the most relevant module names related to the GBO pieces, as shown in Figure 6. The result of this prompt is usually a single module or a list of suggested modules based on the information that the GPT has processed through the chain of prompts. In the next prompt, we included the descriptions of the suggested modules along with the question, as illustrated in the prompt below.

you couldn't give me all the pieces I need in GMO. here is more info that can help: The Funding Award pattern describes the funding awards that fund all kinds of ocean science research activities. We use the isFundedBy property to connect anything to a funding award if the funding award funds it. Each funding award has exactly one starting and ending date (aligned with time:Instant). It provides at most one award amount, which is described via a pair of decimal value and currency code. The currency code is not specified here, but existing standards can be used, e.g., ISO 4217. There may be people or organizations that have a role in a funding award. This is modeled by re-using (and aligning with) the Agent Role pattern. In this version, we include the following types of agent-roles, represented as classes: SponsorRole, AgencyProgramManagerRole, PrincipalInvestigatorRole, and CoPrincipalInvestigatorRole. Additional roles are possible in the future versions. Each funding award is described by an InformationObject, which when aligned to the Information Object pattern, allows one to represent additional information such as identifier, description, etc.

```
ObjectProperty: isFundedBy
ObjectProperty: startsOnDate
ObjectProperty: endsOnDate
ObjectProperty: isDescribedBy
ObjectProperty: providesAgentRole
ObjectProperty: isPerformedBy
ObjectProperty: hasAwardAmount
ObjectProperty: hasCurrencyCode
DataProperty: hasCurrencyValue
Class: FundingAward
Class: Agent
Class: TimeInstant
Class: AgentRole
Class: SponsorRole
Class: PrincipalInvestigatorRole
Class: CoPrincipalInvestigatorRole
Class: AgencyProgramManagerRole
Class: InformationObject
Class: AwardAmount
Class: CurrencyCode
```

Updated Analysis for Equivalence in "GMO"

- **Class "FundingAward"** in GMO corresponds to "Award" in GBO.
- **Object Property "providesAgentRole"** in GMO can be used to connect a "FundingAward" to different roles, including "PrincipalInvestigatorRole" and "CoPrincipalInvestigatorRole".

Constructing the Equivalence Rule

Given the "GBO" relationship:

- **"Award" (x) $\wedge$ "hasCoPrincipalInvestigator" (x, z)**

The equivalent in "GMO", incorporating the additional details provided, would be:

- **"FundingAward" (x) $\wedge$ "providesAgentRole" (x, y) $\wedge$ "CoPrincipalInvestigatorRole" (y) $\wedge$ "isPerformedBy" (y, z)**

Here's the breakdown:

- **"FundingAward" (x):** Directly equates to the "Award" concept in GBO, representing the funding award.
- **"providesAgentRole" (x, y):** Connects a funding award to a role, which can include "CoPrincipalInvestigatorRole".
- **"CoPrincipalInvestigatorRole" (y):** Specifies the role related to the funding award, equivalent to having a co-principal investigator in GBO.
- **"isPerformedBy" (y, z):** Links the role (e.g., CoPrincipalInvestigatorRole) to the person or organization performing it, closing the equivalence to "hasCoPrincipalInvestigator" in GBO.

Fig. 7. GPT-4 response to our question with Module information included

By evaluating the outcome (see Fig. 7) of the prompt, it is evident that GPT-4 has a more informed and precise approach to conducting investigations and identifying semantically related components if it has the module information available. GPT-4 identifies all components related to the GMO as outlined in the alignment rule mentioned above. It further elaborates on each component and explains their interconnections. The full evaluation data of our study over all complex alignment rules is available online.¹⁰

Impact of few-shot vs chain-of-thought: Comparing few-shot and chain-of-thought approaches, we noted the differences between providing information in a single zero-shot prompt versus delivering it in a series of prompts. GPT tends to become confused about the question and the relevant information it needs to process in a zero-shot scenario. In contrast, introducing the prompt

¹⁰ https://github.com/reddraa/complex_ontology_alignment

as part of a sequential chain of information clarifies the data pieces and their meanings for the model.

Impact of adding modules to the prompt: We observed that while GPT-4 nearly grasps the query regarding the relevant components needed to fulfill the rule, it typically identifies only 10-20% of the required elements if no module information is given. However, if details about the module in the GMO are presented in a single prompt before posing the question again, GPT-4 is significantly more effective, clearly identifying the majority of the targeted components.

4.2 Quantitative Evaluation

To assess the effectiveness of our approach, we identified key entities within our complex alignment rules to serve as the basis for our metrics. As illustrated in Figure 7, the entities detected as relevant by GPT-4 include `FundingAwards(z)`, `providesAgentRole(x,y)`, `CoPrincipaInvestigatorRole(y)`, `isPerformedBy(y,z)`. It is important to note that response formats may vary, hence each response might be unique. We employed two key performance metrics: recall and precision. Recall, defined as the ratio of detected GMO-related instances to the total number of expected instances, evaluates our study’s ability to identify all relevant GMO pieces comprehensively. Precision, on the other hand, measures the accuracy of our detection process by calculating the proportion of correctly identified GMO pieces out of all the instances flagged in our findings. Together, these metrics provide a holistic view of our study’s identification capability, striking a balance between thoroughness and accuracy in detecting GMO-related instances of a complex alignment rule. The well-known definitions are as follows.

$$\text{Recall} = \frac{\text{Number of Correctly Detected Pieces}}{\text{Total Number of GMO Pieces in Complex Alignment}}$$

$$\text{Precision} = \frac{\text{Number of Correctly Detected Pieces}}{\text{Total Number of Detected Pieces}}$$

According to Table 1, in the study presented in the paper, an analysis of 109 complex alignment rules was conducted. The findings revealed that in only 4.5% of these alignments, GMO components were identified solely through the use of GBO entities (i.e., not using module information). Both identified complex alignment pieces, without the module, are actually simple 1 : 1 mappings, e.g., `Program(x) ↔ Program(x)`, where ‘Program’ is an existing class in both the GMO and GBO ontology. However, in over 95% of the cases, the identification of GMO components was achieved through the application of information from the GMO module information, highlighting its significant role in the detection process.

In assessing GPT-4’s performance on entity detection in the complex alignment, we analyzed the example alignment mentioned earlier in this paper. We compared the alignment pieces returned by GPT-4, shown in Figure 7 for our

Table 1. Distribution of successful approaches for detecting complex alignment rules

detected pieces without module Information	5
detected pieces with module Information	104
total number of complex alignment rules	109

running example based on rule (1), with four corresponding entities defined in the GMO ontology’s alignment rule, i.e., `FundingAward(x)`, `providesAgentRole(x,y)`, `CoPrincipalInvestigatorRole(y)`, `performedBy(y,z)`. Note that we did not assess the actual return of alignment rules, but rather whether the relevant pieces (predicates) were detected. While the detection (without actual composition of the pieces into a rule) constitutes a simpler task than producing the rules, it nevertheless captures the core difficulty for complex ontology alignment. In fact, if the pieces are correct, the actual rule can easily be assembled by a human (or, in many cases, by a symbolic algorithm based on the ontology and example data).

In evaluating recall (coverage) in this setting, we found that all four expected GMO entities were accurately identified by GPT-4 for our running example, yielding a recall of 1.0 in this case. Additionally, for precision, we examined the entities returned by GPT-4 and found that aside from the four correct GMO entities, no irrelevant entities were detected, also resulting in a precision of 1.0. This indicates perfect alignment detection by GPT-4 in this instance. To further clarify the evaluation of recall, consider that if the prompt response from GPT-4 had included additional entities such as “Event” or “Place” which were not part of the expected entities, it would have negatively impacted the precision. These extraneous entities, not being included in the set of expected results, would reduce the precision as they represent incorrect identifications according to the specified alignment rule. In practical terms, the return of (in particular, many) superfluous entities by the system would make assembly of the actual rule by a human more difficult.

In Table 2 we see that for 73.3% of the complex alignment rules evaluated, the recall value exceeded 0.5. This indicates that more than half of the GMO entities involved in a complex alignment rule were successfully detected for this ratio in our population. Furthermore, the recall value surpasses 0.75 for approximately 62.3% of these rules, signifying a higher accuracy in detection. It’s also noteworthy that *for 45% of the alignment rules recall is a perfect 1.0, and for 45.8% precision is also 1.0*, because of the integration of module information in the analysis process.

The precision metric indicates the accuracy of the responses in directing us toward expected entities within the GMO ontology. Our analysis found that responses achieved a precision higher than 0.75 for 59.6% of the evaluated records. Additionally, when the precision threshold is lowered to 0.5—implying that half of the entities suggested by the language model are the expected ones, while the other half may include relevant or irrelevant entities—the coverage of alignment rules increases to 69.7%.

Table 2. Recall and Precision for detected GMO entities using module information

	Recall			Precision		
	≥ 0.5	≥ 0.75	$= 1$	≥ 0.5	≥ 0.75	$= 1$
with Module Information	73.3%	62.3%	45.0%	69.7%	59.6%	45.8%
mean	0.67			0.67		
median	0.75			0.80		
standard deviation	0.37			0.37		

Note that detection (recall) of half or more of the correct body predicates, paired with high precision, is already very helpful for human assembly of a rule.

The recall data exhibits a mean of 0.67 and a median of 0.75, indicating that half of the recall values exceed 0.75, while the other half fall below this threshold. This distribution highlights that the majority of our data points demonstrate recall values above 0.75, offering insight into the central tendency of our dataset. Similarly, for precision, both the median and mean values suggest promising results, with at least 50% of the records achieving a precision greater than 0.8. This underscores a generally reliable level of accuracy in the data.

We close the discussion with a number of additional more detailed observations.

- Type and Class Type alignments: An intriguing insight from the analysis of complex alignment rules is the difficulty encountered with Type or Class alignment. This process frequently falls short in identifying pertinent alignments. A case in point is the alignment where `GeoFeatureType(x)` in GBO transitions to `rdfs:subClassOf(x, Place)` in GMO. In such instances, the GMO ontology attempts to consolidate individual classes into types or subclasses. Nonetheless, the absence of this specific information in the GMO’s module data impedes the language model’s (LLM) ability to accurately discern these sophisticated alignments, resulting in the provision of erroneous information.
- Rich Modules, Improved Discovery: Another observation is that modules enriched with detailed descriptions, core axioms or alignments, as discussed in the referenced paper [15], depict significantly higher accuracy in discovery. Specifically, modules that offer comprehensive information facilitate better recall and precision rates. For example, the `Cruise()` entity, which involves approximately 18 complex alignment rules, demonstrates an impressive average recall rate of 0.93. This indicates that the depth of information provided directly influences the effectiveness of discovery processes.

5 Discussion and Future Work

The quantitative results we have just presented indicate that our setting—i.e., under inclusion of module information—produces high precision and recall values in many cases. Our results demonstrate a significant improvement compared

to this baseline. While, absolutely speaking, the quantitative results are still moderate, we are in fact presenting the *very first* reasonably working approach for generating complex alignments that does not require shared individuals; as noted, shared individuals are far from a realistic setting in practice. In the latest evaluation of the participating systems in the OAEI complex alignment track, which was in 2021, most systems failed to detect any m:n complex alignments.¹¹ Only two systems procuded complex alignments,¹² but they required instance data. As such our contribution shows a path forward towards complex alignment in realistic settings, a challenge that had so far eluded researchers.

For our approach to work, we provided the LLM with module information, following the previously presented arguments that (1) ontologies without additional internal structure or meaningful additional information are often too ambiguous for automated complex alignment tasks, (2) module identification during the ontology design process can easily be provided by the ontology modelers at least as part of the documentation (while doing so post-hoc, by others, requires major efforts), and (3) providing such module information as part of ontologies (and/or their documentations) would likely significantly decrease the effort and cost of many ontology engineering tasks, including complex alignment [23]. As such, this is also a (repeated) call for improving ontology modeling methods to additionally provide module structure.

While our results, as presented, are very encouraging, substantial future investigations will be required to cast them into an ontology alignment system that can work autonomously at high precision. Intermediate steps could constitute human-in-the-loop approaches where a human ontology engineer receives suggestions from an LLM, e.g., as to the relevant modules for a question, post-processes the LLM responses by manually checking the small number of suggestions, and feeding the correct suggestions back to the LLM for obtaining more complete responses. This is in line with the idea of an assisting system that limits the number of checks for the human, as opposed to the vast number of potential checks that would have to be done manually without such a limiting system.

In the future, we also intend to extend out approach to additional datasets featuring complex alignments for both evaluation and analytical objectives. Furthermore, we aim to explore alternative representations of modules to LLMs and evaluate the model’s performance with these variations. Fine-tuning existing LLMs, but also the integration of additional symbolic data or algorithms, e.g. pertaining to logical axioms that come with well-designed ontologies, and also the integration of traditional simple alignment algorithms to further assist with the complex alignment task are all on our path forward.

6 Conclusion

We have presented the very first approach that is able to achieve good accuracy for complex ontology alignment without relying on shared individuals. The sys-

¹¹ <https://oaei.ontologymatching.org/2021/results/complex/geolink/index.html>

¹² <https://oaei.ontologymatching.org/2021/results/complex/popgeolink/index.html>

tem is neural-symbolic in its nature as it addresses a symbolic task (complex ontology alignment), the output of which are alignment rules expressed in some logic, it furthermore makes decisive use of additional symbolic input in the form of ontology modules, and it uses an LLM as core processing engine. Our results suggest that further work bear the promise to result in strong complex ontology alignment systems, for ontologies that carry sufficient internal structure.

Acknowledgments We acknowledge partial support by NSF award no. 2333532 ”Proto-OKN Theme 3: An Education Gateway for the Proto-OKN”.

References

1. Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
2. Alsayed Algergawy, Michelle Cheatham, Daniel Faria, Alfio Ferrara, Irini Fundulaki, Ian Harrow, Sven Hertling, Ernesto Jiménez-Ruiz, Naouel Karam, Abderrahmane Khiaat, et al. Results of the ontology alignment evaluation initiative 2018. In *13th International Workshop on Ontology Matching co-located with the 17th ISWC (OM 2018)*, volume 2288, pages 76–116, 2018.
3. Alsayed Algergawy, Daniel Faria, Alfio Ferrara, Irini Fundulaki, Ian Harrow, Sven Hertling, Ernesto Jiménez-Ruiz, Naouel Karam, Abderrahmane Khiaat, Patrick Lambrix, Huanyu Li, Stefano Montanelli, Heiko Paulheim, Catia Pesquita, Tzanina Saveta, Pavel Shvaiko, Andrea Splendiani, Élodie Thiéblin, Cássia Trojahn, Jana Vataschinová, Ondrej Zamazal, and Lu Zhou. Results of the ontology alignment evaluation initiative 2019. In Pavel Shvaiko, Jérôme Euzénat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn, editors, *Proceedings of the 14th International Workshop on Ontology Matching co-located with the 18th International Semantic Web Conference (ISWC 2019), Auckland, New Zealand, October 26, 2019*, volume 2536 of *CEUR Workshop Proceedings*, pages 46–85. CEUR-WS.org, 2019.
4. Reihaneh Amini, Lu Zhou, and Pascal Hitzler. Geolink cruises: A non-synthetic benchmark for co-reference resolution on knowledge graphs. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2959–2966, 2020.
5. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
6. Michelle Cheatham and Pascal Hitzler. Conference v2. 0: An uncertain version of the oaei conference benchmark. In *The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II 13*, pages 33–48. Springer, 2014.
7. Michelle Cheatham, Adila Krisnadhi, Reihaneh Amini, Pascal Hitzler, Krzysztof Janowicz, Adam Shepherd, Tom Narock, Matt Jones, and Peng Ji. The geolink knowledge graph. *Big Earth Data*, 2(2):131–143, 2018.
8. Jérôme Euzénat and Pavel Shvaiko. *Ontology Matching, Second Edition*. Springer, 2013.

9. Yuan He, Jiaoyan Chen, Hang Dong, and Ian Horrocks. Exploring large language models for ontology alignment. In Irini Fundulaki, Kouji Kozaki, Daniel Garijo, and José Manuel Gómez-Pérez, editors, *Proceedings of the ISWC 2023 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 22nd International Semantic Web Conference (ISWC 2023)*, Athens, Greece, November 6-10, 2023, volume 3632 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2023.
10. Sven Hertling and Heiko Paulheim. Olala: Ontology matching with large language models. In *Proceedings of the 12th Knowledge Capture Conference 2023*, pages 131–139, 2023.
11. Pascal Hitzler. A review of the semantic web field. *Commun. ACM*, 64(2):76–83, 2021.
12. Pascal Hitzler, Markus Krötzsch, and Sebastian Rudolph. *Foundations of Semantic Web Technologies*. Chapman and Hall/CRC Press, 2010.
13. Pascal Hitzler, Md. Kamruzzaman Sarker, and Aaron Eberhart, editors. *Compendium of Neurosymbolic Artificial Intelligence*, volume 369 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2023.
14. Adila Krisnadhi, Yingjie Hu, Robert Arko, Suzanne Carbotte, Cynthia Chandler, Michelle Cheatham, Douglas Fils, Timothy Finin, Pascal Hitzler, Krzysztof Janowicz, et al. Geolink core ontology design patterns. Technical report. Available from <https://people.cs.ksu.edu/~hitzler/pub2/gmo-tr.pdf>.
15. Adila Krisnadhi, Yingjie Hu, Krzysztof Janowicz, Pascal Hitzler, Robert Arko, Suzanne Carbotte, Cynthia Chandler, Michelle Cheatham, Douglas Fils, Timothy Finin, et al. 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 14*, pages 301–309. Springer, 2015.
16. Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.
17. Sanaz Saki Norouzi, Mohammad Saeid Mahdavejad, and Pascal Hitzler. Conversational ontology alignment with chatgpt. In Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn, editors, *Proceedings of the 18th International Workshop on Ontology Matching co-located with the 22nd International Semantic Web Conference (ISWC 2023)*, Athens, Greece, November 7, 2023, volume 3591 of *CEUR Workshop Proceedings*, pages 61–66. CEUR-WS.org, 2023.
18. Mina Abd Nikooie Pour, Alsayed Algergawy, Florence Amardeilh, Reihaneh Amini, Omaira Fallatah, Daniel Faria, Irini Fundulaki, Ian Harrow, Sven Hertling, Pascal Hitzler, Martin Huschka, Liliana Ibanescu, Ernesto Jiménez-Ruiz, Naouel Karam, Amir Laadhar, Patrick Lambrix, Huanyu Li, Ying Li, Franck Michel, Engy Nasr, Heiko Paulheim, Catia Pesquita, Jan Portisch, Catherine Roussey, Tzanina Saveta, Pavel Shvaiko, Andrea Splendiani, Cássia Trojahn, Jana Vataschinová, Beyza Yaman, Ondrej Zamazal, and Lu Zhou. Results of the ontology alignment evaluation initiative 2021. In Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn, editors, *Proceedings of the 16th International Workshop on Ontology Matching co-located with the 20th International Semantic Web Conference (ISWC 2021)*, Virtual conference, October 25, 2021, volume 3063 of *CEUR Workshop Proceedings*, pages 62–108. CEUR-WS.org, 2021.
19. Mina Abd Nikooie Pour, Alsayed Algergawy, Reihaneh Amini, Daniel Faria, Irini Fundulaki, Ian Harrow, Sven Hertling, Ernesto Jiménez-Ruiz, Clément Jonquet,

- Naouel Karam, Abderrahmane Khiat, Amir Laadhar, Patrick Lambrix, Huanyu Li, Ying Li, Pascal Hitzler, Heiko Paulheim, Catia Pesquita, Tzanina Saveta, Pavel Shvaiko, Andrea Splendiani, Élodie Thiéblin, Cássia Trojahn, Jana Vataschinová, Beyza Yaman, Ondrej Zamazal, and Lu Zhou. Results of the ontology alignment evaluation initiative 2020. In Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn, editors, *Proceedings of the 15th International Workshop on Ontology Matching co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual conference (originally planned to be in Athens, Greece), November 2, 2020*, volume 2788 of *CEUR Workshop Proceedings*, pages 92–138. CEUR-WS.org, 2020.
20. Mina Abd Nikooie Pour, Alsayed Algergawy, Patrice Buche, Leyla Jael Castro, Jaoyan Chen, Adrien Coulet, Julien Cufi, Hang Dong, Omaina Fallatah, Daniel Faria, Irini Fundulaki, Sven Hertling, Yuan He, Ian Horrocks, Martin Huschka, Liliana Ibanescu, Sarika Jain, Ernesto Jiménez-Ruiz, Naouel Karam, Patrick Lambrix, Huanyu Li, Ying Li, Pierre Monnin, Engy Nasr, Heiko Paulheim, Catia Pesquita, Tzanina Saveta, Pavel Shvaiko, Guilherme Sousa, Cássia Trojahn, Jana Vataschinová, Mingfang Wu, Beyza Yaman, Ondrej Zamazal, and Lu Zhou. Results of the ontology alignment evaluation initiative 2023. In Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn, editors, *Proceedings of the 18th International Workshop on Ontology Matching co-located with the 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 7, 2023*, volume 3591 of *CEUR Workshop Proceedings*, pages 97–139. CEUR-WS.org, 2023.
 21. Zhangcheng Qiang, Weiqing Wang, and Kerry Taylor. Agent-om: Leveraging large language models for ontology matching. *CoRR*, abs/2312.00326, 2023.
 22. Sebastian Rudolph, Markus Krötzsch, Peter Patel-Schneider, Pascal Hitzler, and Bijan Parsia. OWL 2 web ontology language primer (second edition). W3C recommendation, W3C, December 2012. <https://www.w3.org/TR/2012/REC-owl2-primer-20121211/>.
 23. Cogan Shimizu, Karl Hammar, and Pascal Hitzler. Modular ontology modeling. *Semantic Web*, 14(3):459–489, 2023.
 24. Pavel Shvaiko and Jérôme Euzenat. Ontology matching: state of the art and future challenges. *IEEE Transactions on knowledge and data engineering*, 25(1):158–176, 2011.
 25. Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
 26. Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.
 27. Lu Zhou, Michelle Cheatham, Adila Krisnadhi, and Pascal Hitzler. A complex alignment benchmark: Geolink dataset. In *The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II 17*, pages 273–288. Springer, 2018.
 28. Lu Zhou, Michelle Cheatham, Adila Krisnadhi, and Pascal Hitzler. Geolink data set: A complex alignment benchmark from real-world ontology. *Data Intelligence*, 2(3):353–378, 2020.



The GeoLink knowledge graph

Michelle Cheatham, Adila Krisnadhi, Reihaneh Amini, Pascal Hitzler, Krzysztof Janowicz, Adam Shepherd, Tom Narock, Matt Jones & Peng Ji

To cite this article: Michelle Cheatham, Adila Krisnadhi, Reihaneh Amini, Pascal Hitzler, Krzysztof Janowicz, Adam Shepherd, Tom Narock, Matt Jones & Peng Ji (2018) The GeoLink knowledge graph, Big Earth Data, 2:2, 131-143, DOI: [10.1080/20964471.2018.1469291](https://doi.org/10.1080/20964471.2018.1469291)

To link to this article: <https://doi.org/10.1080/20964471.2018.1469291>



© 2018 The Author(s). Published by Taylor & Francis Group and Science Press on behalf of the International Society for Digital Earth, supported by the CAS Earth Strategic Priority Research Programme.



Published online: 18 May 2018.



Submit your article to this journal [↗](#)



Article views: 4820



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 8 View citing articles [↗](#)



DATA ARTICLE

OPEN ACCESS



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
and Peng Ji^g

^aDepartment of Computer Science and Engineering, Wright State University, Dayton, OH, USA;

^bDepartment of Computer Science, Universitas Indonesia, Depok, Indonesia; ^cGeography Department, University of California at Santa Barbara, Santa Barbara, CA, USA; ^dWoods Hole Oceanographic Institution, Woods Hole, MA, USA; ^eMathematics/Physics Department, Notre Dame of Maryland University, Baltimore, MD, USA; ^fNational Center for Ecological Analysis and Synthesis, Santa Barbara, CA, USA; ^gLamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA

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 Protocol and RDF Query Language (SPARQL) end point that at present contains more than 45 million Resource Description Framework (RDF) triples together with a collection of ontologies and geo-visualization tools. This article 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>.

ARTICLE HISTORY

Received 13 February 2018

Accepted 19 April 2018

KEYWORDS

Geoscience data;
oceanographic data;
knowledge graph; linked
data; 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 to data

CONTACT Michelle Cheatham  michelle.cheatham@wright.edu

© 2018 The Author(s). Published by Taylor & Francis Group and Science Press on behalf of the International Society for Digital Earth, supported by the CAS Earth Strategic Priority Research Programme.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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 article, we present the GeoLink knowledge graph, from the data it contains, to how it is modeled and 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 Program⁴ 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) Library⁵ 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 Earth⁷ 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 Database⁸ 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 Program⁹ 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), articles 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 Resource Description Framework (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 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).

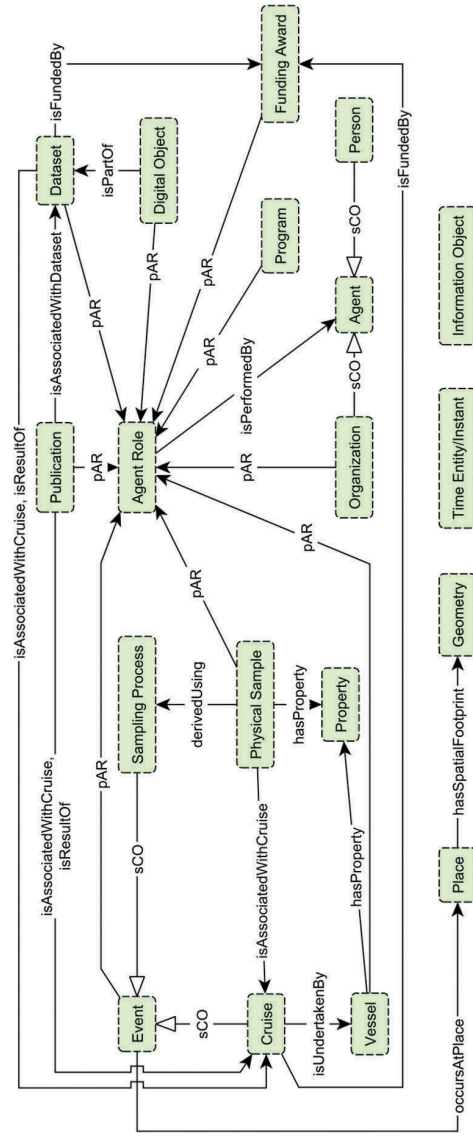


Figure 1. Schema diagram containing (almost) all patterns in the Geolink ontology and their main links.

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.

3.2. The GeoLink base ontology

While the GMO follows best-practices in ontological modeling (Hitzler, Gangemi, Janowicz, Krisnadhi, & 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 Uniform Resource Identifiers (URIs).

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.

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 Protocol and RDF Query Language (SPARQL) construct queries can then be used to convert the information into the

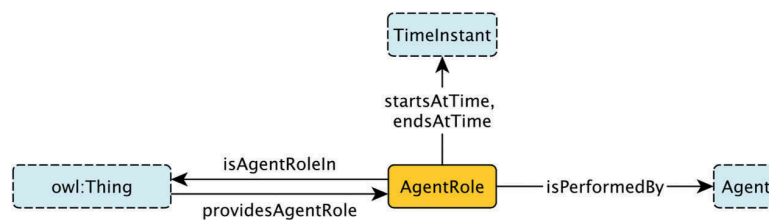


Figure 2. The agent role pattern.



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

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's¹⁰ 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 Friend of a Friend (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 Hyper Text Transfer Protocol (HTTP) end point, 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 and title and publication date for articles). 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 (ORCIDiDs) and Digital Object Identifiers (DOIs). ORCIDiDs 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. ORCIDiDs 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 (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: Nonproprietary 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, nonproprietary, 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 available¹⁶ 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).¹⁷ 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 toward 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 System¹⁹ 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 from the NOAA, National Centers for Environmental Information, the Index to Marine and Lacustrine Geological Samples, and the CLIVAR & Carbon Hydrographic Data Office 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 3-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

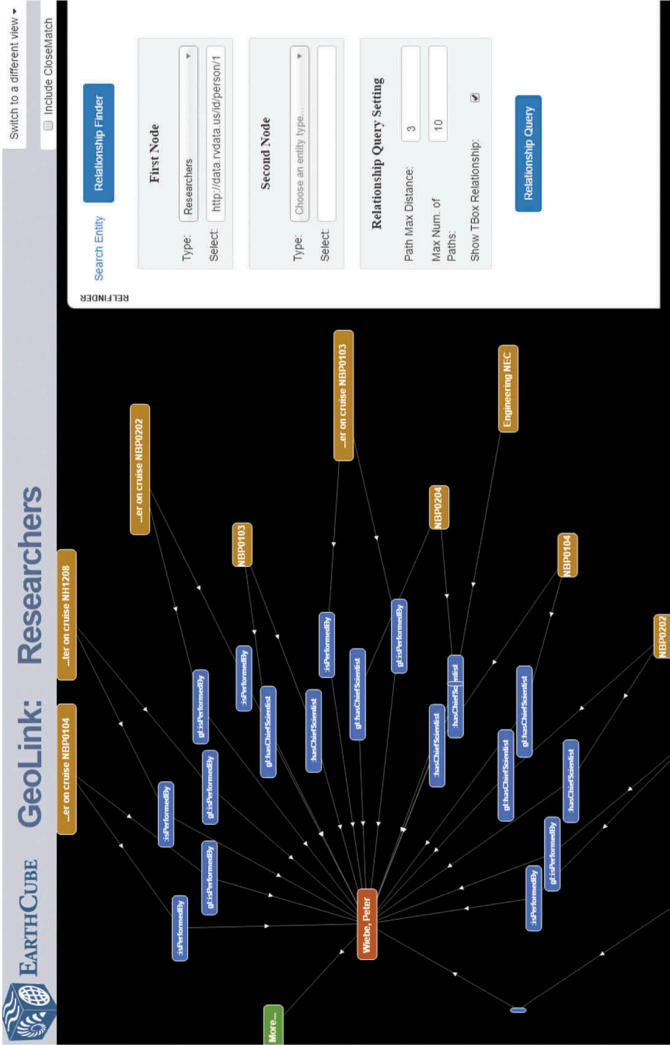


Figure 4. A web-based tool that allows users to explore the GeoLink linked data set.

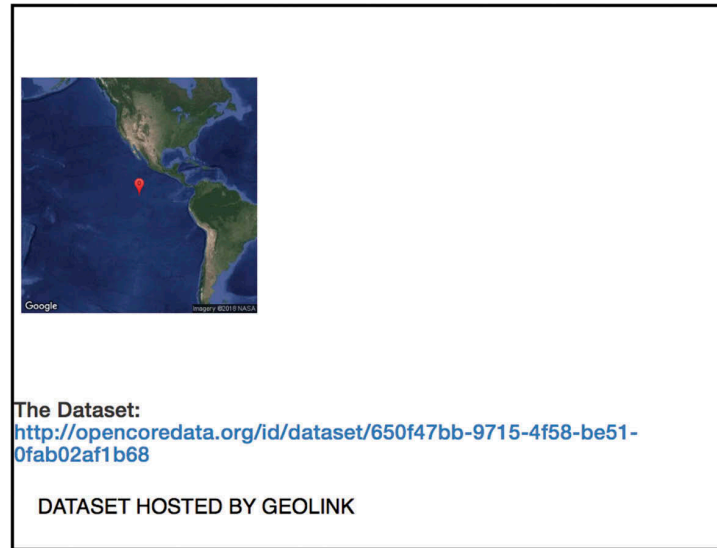


Figure 5. A GeoLink web component, which allows web developers to easily incorporate GeoLink data onto their sites.

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 one-to-one 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 datasets are more complicated, 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 one-to-one 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 (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 aligning 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 end point is maintained by the National Center for Ecological Analysis and Synthesis (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 article 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. 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.

Data availability statement

The data described in this article are openly available in the MBLWHOI Library at <http://doi.org/10.1575/1912/9524>.

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/Data>
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>

Acknowledgments

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.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Science Foundation [1440202].

ORCID

Adila Krisnadhi  <http://orcid.org/0000-0003-0745-6804>
 Adam Shepherd  <http://orcid.org/0000-0003-4486-9448>
 Tom Narock  <http://orcid.org/0000-0002-9785-4496>
 Matt Jones  <http://orcid.org/0000-0003-0077-4738>
 Peng Ji  <http://orcid.org/0000-0003-1868-5004>

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). Aachen. [ceur-ws.org](http://www.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). Amsterdam: IOS Press/AKA Verlag.
- Krisnadhi, A., Hu, Y., Janowicz, K., Hitzler, P., Arko, R., Carbotte, S., ... Weibe, P. (2015a). The GeoLink framework for pattern-based linked data integration. In *Proceedings of the posters and demos session at the 14th International Semantic Web Conference* (Vol. 1486). Aachen. [ceur-ws.org](http://www.ceur-ws.org)
- Krisnadhi, A., Hu, Y., Janowicz, K., Hitzler, P., Arko, R., Carbotte, S., ... Weibe, P. (2015b). The GeoLink modular oceanography ontology. In *Proceedings of the 14th International Semantic Web Conference* (Vol. 9367, pp. 301–309). Cham: Springer.