2023-05-01

Enhancing Basic Geology Skills With Artificial Intelligence: An Exploration Of Automated Reasoning In Field Geology

Perry Ivan Quinto Houser
University of Texas at El Paso

ENHANCING BASIC GEOLOGY SKILLS WITH ARTIFICIAL INTELLIGENCE:

AN EXPLORATION OF AUTOMATED REASONING IN FIELD GEOLOGY

PERRY IVAN QUINTO HOUSER

Master's Program in Earth, Environmental and Resource Sciences

APPROVED:

Deana D. Pennington, Ph.D., Chair

James D. Kubicki, Ph.D., Co-Chair

Jason W. Ricketts, Ph.D.

Natalia Villanueva Rosales, Ph.D.

Stephen L. Crites, Jr., Ph.D.
Dean of the Graduate School

Dedication

To the memorable and critical moment when the study of geology enveloped me with the unique experience of connecting my love of nature, exploration, and curiosity with my love of technology and visualization that represents the core of this thesis. My journey to this point would not have been possible without the invaluable experiences and support of my family, mentors, and friends. Their encouragement ignited my passion to pursue this endeavor with an open mind, kindness, and gratitude. I am forever grateful for their unwavering guidance and inspiration. Completing this thesis has allowed me to come full circle, and I hope that in some way, I have inspired others on their own quest for knowledge and self-discovery. Moving forward, I am committed to continuing to do my best in making a positive impact for myself and on those around me.

Thank you.

ENHANCING BASIC GEOLOGY SKILLS WITH ARTIFICIAL INTELLIGENCE: AN EXPLORATION OF AUTOMATED REASONING IN FIELD GEOLOGY

by

PERRY IVAN QUINTO HOUSER, BACHELORS IN GEOLOGY

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Earth, Environmental and Resource Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

May 2023

Acknowledgements

To a whole team of people that supported with me on this journey: my wonderful wife, my children; my father and mother; my extended family members and in-laws, and to my grandmother in-law, who always reminded me that perseverance is key.

Thank you for the mentorship of my advisors: Dr. Deana D. Pennington, and Dr. James D. Kubicki; and my committee members: Dr. Jason W. Ricketts, and Dr. Natalia Villanueva Rosales. Thank you to Dr. Elizabeth Walsh (Biology), who is an inspiration and invaluable support for myself and as my wife's advisor. Forever grateful to my undergraduate advisor: Dr. Bridget Smith-Konter; and so many professors along way: Dr. Jose M. Hurtado, Dr. Richard Langford, Dr. Philip Goodell, Dr. Thomas Gill, Dr. Benjamin Brunner, Dr. Gale Arnold, Dr. Terry Pavlis, Dr. Laura Serpa, and Dr. Aaron A. Velasco – who introduced me to geology. Thank you to Dr. Clarence H. Cooper (Physics); and so many more that molded me to who I am now; the Geology department, the skilled and caring support provided by the Graduate Coordinator, Dr. Veilleux; and the other key departments and faculty and staff within the university. Thank you to Dr. Yolanda Gil (USC) for the outside world experience and her skilled group of scientists during my internship. Lastly, thanks to the many friends and colleagues that also fostered my vision, knowledge, and personal growth during my academic pursuits.

This material is based upon work supported by the National Science Foundation under Grant No. HRD-1242122 for the Cyber-ShARE Center of Excellence. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Abstract

This thesis explores the use of Artificial Intelligence, specifically semantics, ontologies, and reasoner techniques, to improve field geology mapping. The thesis focuses on two use cases: 1) identifying a geologic formation based on observed characteristics; and 2) predicting the geologic formation that might be expected next based upon known stratigraphic sequence. The results show that the ontology was able to correctly identify the geologic formation for the majority of rock descriptions, with higher search results for descriptions that provided more detail. Similarly, the units expected next were correctly given and if incorrect, would provide a flag to the field geologist to further investigate the sequence break. However, subjective descriptions and searches can impact the results, and incorrect property assertions can generate undesirable results and require validation and verification of data. Overall, the study demonstrates the potential for using sematic knowledge bases for field studies to improve geologic field observations and measurements.

Table of Contents

List of Tables

List of Figures

Introduction

A key component of geoscience research and education is field work, which is typically accomplished by recording data into geospatial mapping platforms and archiving observations into digital or written notebooks and plotted on maps. These approaches generally do not take advantage of the technical strengths that cyberinfrastructure (CI) systems have to offer, which provide opportunities to digitally render and preserve the fullest intent and proficiency of a field geologist. Hey and Trefethen (2005) describe CI as a collective system of computer hardware, software, networks, data, semantics, models, and human collaborators, often based on cutting edge technology. Tim Berners-Lee's initial conception of the World Wide Web (WWW) envisioned an online environment for the scientific community to process and analyze information that goes beyond the capabilities of human processing alone (Berners-Lee et al., 2001). A WWW that combines human knowledge with machine reasoning about information, termed the "semantic web," has been the subject of decades of research and is beginning to come to fruition as a component of emerging intelligent systems (IS), along with new machine learning techniques.

Continual advancements in CI presents opportunities for improvement within the earth science community including analyzing big data sets, developing novel workflows, and creating tools that analyze the information collected (Pennington et al., 2020; Plale et al., 2013). The education section of the American Geophysical Union (AGU) recently concluded that the changing landscape of information technology (e.g., big data, tools, models, collaborators) affects the kinds and quantities of resources that are available for problem solving. Both students and domain experts must learn to navigate this rapidly changing space by successfully identifying and harnessing resources that can be brought to bear (Brown et al., 2008; Christensen & Knezek, 2019). Hey & Trefethen (2005) state, "In order to exploit and explore" the sheer amount of information

present, a degree of automation is required and can be provided by the semantic web elements within cyberinfrastructure that process data and metadata. Semantic web formats these data types into an environment appropriate for computer-to-computer derived logic (machine reasoning) while still being comprehensible to humans.

These new and emerging technologies are well-suited for traditional geological field mapping. Geologic field mapping is a fundamental principle of data gathering and verification; first-hand field data is considered the most thorough and authoritative method of collecting geologic observations and measurements (Swetanisha, 2022). Geological field mapping exercises can and should be enabled and enhanced by a variety of emerging technologies (Neumann et al., 2006; Mookerjee et al., 2014, Scianna et al., 2012; Sinha et al., 2010). The acceptance of newer technologies among geoscientists continues to grow, while not replacing analog data but instead being integrated in parallel with traditional non-digital methodologies of using pen and paper maps, sketches, journaling, and even abstract hand gestures to understand complex 3D structures and motion (Lundmark, 2020).

Early introduction of students to field mapping is gaining popularity with professors, as a study by Elkins (2007) demonstrated higher geoscience cognitive improvement and increased geologic interest for students that participated in field research over those that completed analogous studies in a classroom setting. Recording and understanding in-situ data is considered a critical part of the geologic field training (Compton, 1985). New methods for collecting and analyzing data powered by machine learning techniques provide semi-automatic, novel methods of data management workflows and enhances discovery in the field. The National Science Foundation's EarthCube program supports community driven research on new approaches to enable cyberinfrastructure geared towards geospatially referenced field and microstructural

observations; integrate basic analyses; visualize interpreted structural histories; integrate geophysical data (geodetic, seismological); semi-automate digitizing; and provide digital laboratory and field notebooks (Smith et al., 2016).

One way to aid traditional field data collection is using semantic-based technologies. Semantic-based technologies are a mechanism for representing knowledge (e.g., internal mental models of a topic) in machine-readable form that enables machine reasoning and inferencing of information that is not explicitly entered into a database by a human user (Fonseca et al., 2002; Gil et al., 2016). Both humans and machines need some sort of context, or semantics, to impart meaning on a piece of data (Mitchell, 2019). For example, the word "age" could refer to a person's or object's age, the process of aging, or an historical era. The meaning of the word is inferred from the context, or semantics, of its usage. Semantic web knowledge bases aim to improve search results by representing concepts derived from keywords used to tag data (vocabulary) and the context (semantics) of their usage through the development and use of "ontologies" that describe concepts and relationships in machine-readable formats, specifically in the Web Ontology Language or OWL (Parsia, 2012). This differs from standard data models in that ontologies enable machine reasoning and inferencing driven by observations that are recorded to the knowledge base. Noy (2001) describes ontologies as a formalized description of a vocabulary (concepts) within a specific domain. An ontology is a model of the concepts and relationships that are relevant in a particular context. Like any model, choices are made by the developers about how to simplify and organize the phenomena being rendered, what concepts should be represented in the model and how they are related. Different contexts and varying levels of expertise of the ontology developers invariably result in distinct descriptions of terminology, relationships, and structure (i.e., mental models and workflow). Intelligent systems can bridge the gap between disparate ontologies

constructed in an "open-world" framework via the re-use of established lexicons, and repositories of community-driven ontologies that strengthen machine learning and reasoning (Gil, et. al, 2016). Semantically annotated data (instances) together with their ontologies are referred to as a semantic web knowledge base. Ontologies enable machine reasoning that goes beyond standard searches of simple string-matching (syntactic) methods (Kasenchak, 2019). The goal of machine inferencing is to connect and integrate data that may be from disparate data sources and/or be logically organized in different ways into a single cohesive environment for use by intelligent systems.

The National Science Foundation's EarthCube program has established a foundational, multi-level support system for effective representation and analysis of earth data within an intelligent, knowledge rich platform (Gil et al., 2018). EarthCube's Macrostrat offers a general data model to test geologic hypotheses regarding rock preservation and cycling along with biologic drivers based off microstratigraphy mapping research from Peters (2018). Macrostrat uses a relational database and *Structured Query Language* (SQL) driven quantitative analysis to provide big data scale synthesis of relatively low-resolution regional geologic column field data to high-resolution local primary field rock unit observations and measurements that assist with a more complete high-level descriptions of the Earth's upper crust (Peters et al., 2018). EarthCube's StraboSpot covers structural geology applications within a cyberinfrastructure environment relating field data and laboratory analysis using a graph database and NOSQL (*not only SQL)* analysis instead of a SQL based relational database based on model data as records in rows and tables with logical links between them (Walker et al., 2019). Both EarthCube platforms provide elements of data management, but they do not incorporate machine reasoning and inferencing as envisioned by semantic and ontology approaches.

This thesis focuses on the application of formal semantics and ontological techniques to field geology. The overarching question of interest is: How can these techniques support training geoscientists in the field? I approach this by analyzing two use cases:

1) Identifying a geologic formation in the field based on observed characteristics of the rock outcrops; and
2) Predicting the appearance of a formation in the field given the observed geologic formation and the known stratigraphic sequence in the area.

Background

GEOLOGIC FIELD MAPPING

Geologic field mapping offers a unique opportunity for geologists to directly interact with, observe, measure, and interpret the environment or phenomena in its natural state (Priestnall et al., 2007). Basic geospatial cognitive abilities involve gathering and understanding information of 3D structures paired with possible dynamic geologic processes occurring over long periods of time (Saini-Eidukat, et. al, 2002).

Key principles of field geology traditionally include basic procedures undertaken at outcrops, which involve proper identification and measurements of geologic structures and rocks found. Other key principles involve the proper use of equipment for sampling and recording observations, including the use of a compass and maps (geologic maps and topographic maps), along with aerial photographs or remote sensing imagery, if available. Since field geology requires interpretation of land, it is common to have illustrations of a specific feature to convey information that does not have to be duplicated in the final report of the area, and/or to visually assist the report with information that may not be apparent to the reader. Further aids come from stratigraphic sections that define the sequences of rock, geologic structure, and relative position that dictate ages and possible events which occurred to deposit, deform, or alter the conditions on the ground (Compton, 1985). The knowledge base introduction as a field tool would be best served for a couple of audiences. One, geoscience students who are already familiar with the foundational practices of geologic field mapping and are using the knowledge base as a scientific logical aid developing the ontology for data capture and discovery with the purpose of building a story of the geology while still in the field. As Compton (1985) states, it is critical to "recognize key features the first-time around" as the mapper may only have a single opportunity to visit the

site. Another target audience would be field studies that involve “experts” in the field verifying past or co-existing data in the field. Compton states that collecting, recording, and interpreting geological information, both user-created and officially recognized information recorded such as proper names, symbols, and colors should be accurate and use standards set forth by organizations such as the United States Geological Survey (USGS) Domestic Names Committee of the U.S. Board on Geographic Names. Ontologies extends standardization conventions provided by a series of possible organizations and concepts, such as the International Organization for Standardization (ISO, 2021), World Wide Web Consortium (W3C, 2023, Open Geospatial Consortium (OGC, 2021), Spatial Data Infrastructures (Janssen et al., 2012), National Information Standards Organization (NISO, 2021), and the American National Standards Institute (ANSI, 2021). The intent of this thesis study is to introduce supplementary semantic tools into the field mapping workflow and assess their application in this context. Capabilities of identifying rocks and mineral, structure and processes should be established as a foundational cogitative learning process. Semantic web knowledge bases should act to foster organization, sharing, and building understanding in real-time while in the field, even if the stratigraphic column and units that belong to it are being discovered and named. The queries can be applied generically to any site if the competency questions and ontology structure are addressed first.

SEMANTIC WEB KNOWLEDGE BASES

Encapsulating information into Semantic Web Knowledge Bases (SWKB) depend on integrating three perspectives: 1) ontology engineer; 2) domain expert; and 3) end user (Kalbasi 2014). The end user defines the problem and context and articulates the specific questions the ontology is intended to address. The domain expert has a mental model of the concepts and

relationships that are relevant to that problem. The ontology engineer develops the technology that captures the concepts and relationships using a tagged structure of terminology, relationships, and constraints in ways that can address the problem of interest. Information about where the data came from is kept within the provenance data provided within Protégé or can be imported as well via a provenance ontology such as PROV-N (Lebo et al., 2013).

The Protégé Ontology Editor (Musen, 2015) uses “Class”, “Object/Data Properties”, “Instances” and “Annotations” features. All data encoded into the SWKB must be designated as one of these. The foundational structure consists of classes and subclasses that represent relevant concepts in a hierarchical vocabulary. These are linked through properties that describe the relationship between (sub)classes. Instances are specific cases of a (sub)class that do not require further subclassing. Annotations are applied to classes, properties, and instances to capture detailed information about each element, along with annotations about the ontology itself, such as version tracking and general comments.

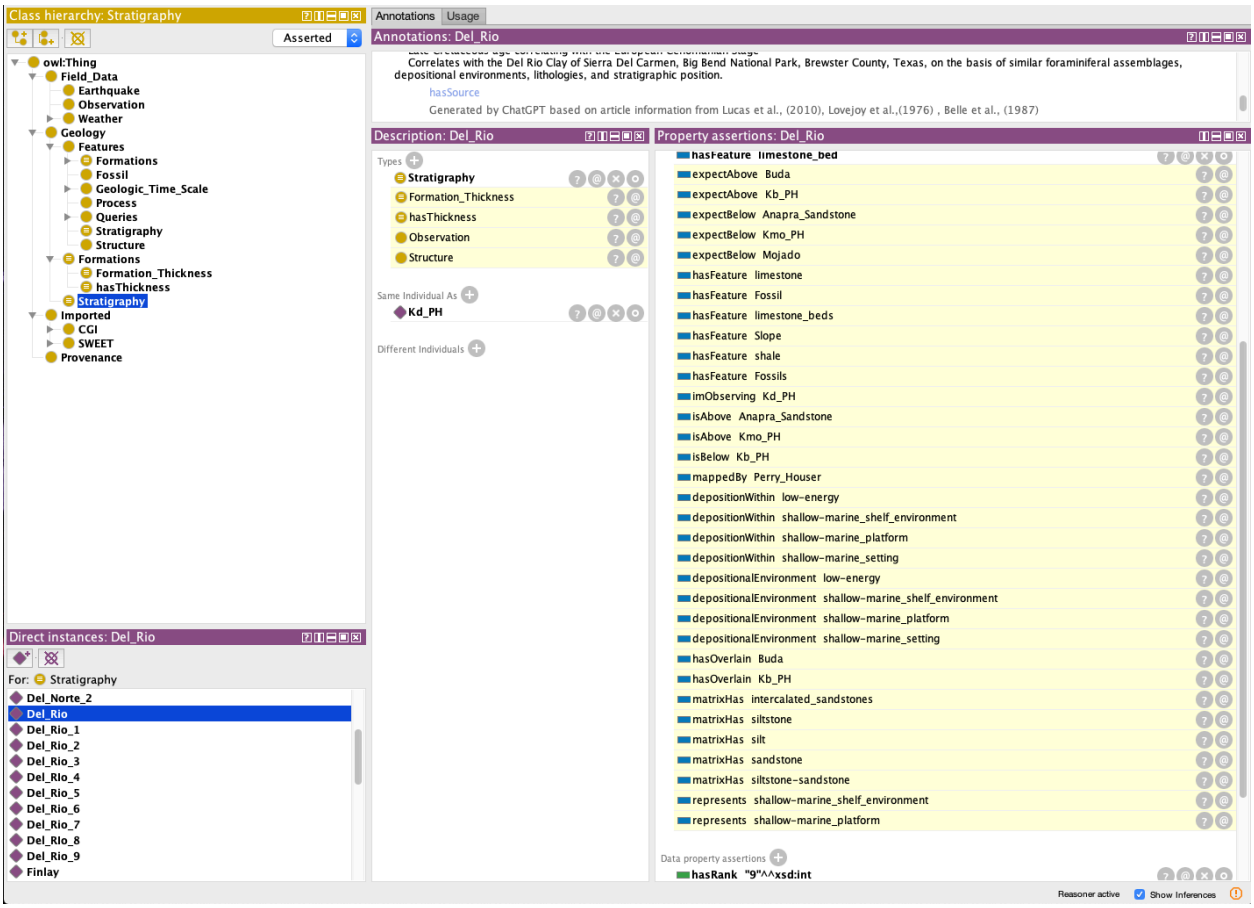

Figure 1: Example of ontology structure and inferences using Protégé version 5.5.0. A role-chain driven inference is shown in yellow is based on what one might expect next as a property assertion: “isAbove” and “expectBelow”, are determined by which unit being observed “Del_Rio” Formation.

Programs such as Protégé provide a user-friendly front-end application for ontology development and management. Once developed, the ontology must be tested to validate that it is able to support the end user in the way envisioned. This process requires the development of ontology “competency questions” that articulate what inferences are needed and enable testing against data sets where the expected results are known. Inferences are derived by the reasoner to generate implicit information that is not explicitly entered into the ontology by the user, using a set of defined rules and logical axioms. Geologic knowledge and field data are organized into data classes, subclasses, and relationships or may be sourced from pre-existing ontologies.

Competency questions can be answered by simple searches of keywords and/or more complicated queries using a Description Logic Query. The DL Query tab in Protégé provides more advanced and/or alternative methods for searching and included input from the reasoner in the summary of supported results for a competency question specific to formal knowledge representation. Protégé offers the ability to represent description logics with a user-friendly syntax – Manchester syntax (Smith et al., 2021). Using Manchester Syntax, Description Logics classes can be defined and with the use of a reasoner, instances, subclasses, superclasses and other information can be retrieved using the DL Query tab. These DL classes can answer competency questions. These requires a syntactically correct formatted statements based upon familiarly with the basics of ontological concepts, roles, individuals, and axioms. Protégé facilitates construction of the statements since they are more sophisticated than simple text searches. For example, Protégé can list properties for a geologic formation is listed as both explicitly entered and inferences implicitly derived by the reasoner (Figure 1). Implicit data is shown in yellow, includes an example of enabling a role-chain within the object property "expectAbove" using the "SuperProperty Of (Chain)" shown in Figure 2.

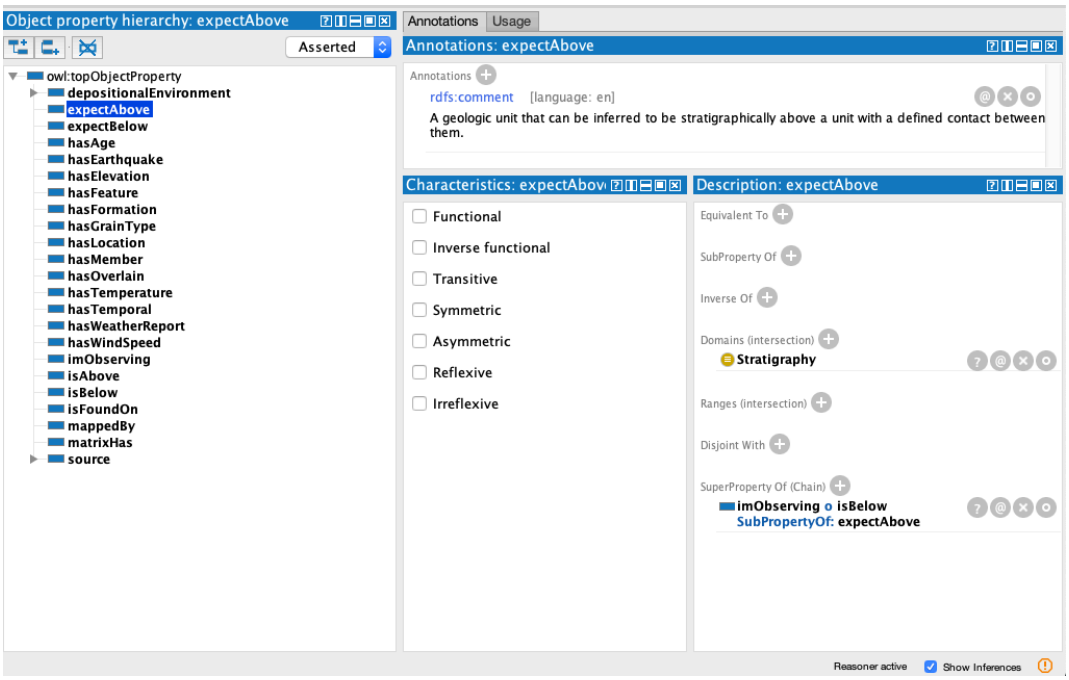

Figure 2: Using a role-chain for inferences on what geologically might be expected next in Protégé version 5.5.0.

Ontologies are developed to support a specific problem and tend to be partially reused along with more broadly scoped ontologies (upper-level ontologies) that have been established using standardized lexicons from expert domains. Reuse of broadly scoped ontologies by multiple problem-specific ontologies provides a means of integrating data across problem contexts. This means that SWKB developers must not only generate tailored ontologies that represent the mental models of a domain expert for a limited problem, but also consolidate existing ontologies that were developed for their own specific research questions and scope.

In the context of geological field mapping, ontologies can supplement a field notebook and/or provide another field mapping partner as a digital extension to aid with data capture, sharing, and geologic hypothesis development. The functionality of the ontology is based on what information will be held and what questions are to be answered. Commonly collected observations in geological field mapping are shown in Table 1, along with existing ontologies

that include those concepts. The main functions that must be covered by a geologic field mapping ontology include: 1) Provenance information; 2) Meta-data; 3) Time and date; 4) Weather; 5) Location and position; along with 6) Domain specific records (geologic data). The first five of these are common to all field science problems. The sixth, domain specific records, must be developed for the specific problem of interest.

Table 1. Generalized information covered by both geologic field notebook and field ontologies.

Category	Examples	Ontology Coverage Source(s) IRIs
Provenance	Entity information: author name(s), membership, role; Activity: project title, purpose; Time & Date: start, end; Indentifier: version, revision; Bundle: single, collection, (provenance of provenance data); Alternate information: specialization, confidence	PROV-N: The Provenance Notation, http://purl.dataone.org/provone/2015/01/15/ontology#
Metadata	Domain relevant: annotations, descriptions, comments, labels, and/or identifiers	Protégé ontology annotation property hierarchy
Imported field data	Local weather and earthquake data	NOAA Weather Ontology, USGS QuakeML https://api.weather.gov/ontology, https://quake.ethz.ch/quakeml/QuakeML2.0
Location & Position	Geospatial features: spatial relationships (i.e., next-to), toponyms (place names), coordinate reference systems, grids, metadata, services	W3C Geospatial Ontologies, https://www.w3.org/2005/Incubator/geo/XGR-geo-ont-20071023/
Geologic Observations & Measurements	Expert domains: stratagraphy, geologic timescale, paleobiology, geochemistry, earthquakes, hydrology, unit data	NASA SWEET Ontology, CGI http://sweetontology.net/sweetAll, https://cgi.vocabs.ga.gov.au/vocab/

For example, a common geologic field observation is a fault that cuts exposed rock units. When recording the description of a fault, basic information such as the name of the fault (if known), location, relationship to geologic units, and other possible measurements are required such as movement sense. The SWKB developer's use of variables within the ontology may have to take into consideration domain information such as relationships to outcrops, geologic time periods, and specific geologic processes. It may be necessary to specify technical information

such as data types (e.g., a String, Float, Integer) and units. It will also ideally include metadata about the record. Provenance, such as who recorded the information, relationships to institutions, language (e.g., English), and when the information was recorded may have its own metadata, yet metadata covers more than just provenance. Complexity arises as there are many correct methods for achieving this. One method might be to solely rely upon a simplified user-made variable names for each required entry to encompass the properties of the fault; another, more desirable method, would be to use an existing set of variable names already in use by a trusted institution. Table 2 provides a comparison of the approach used by two different ontologies to describe a geologic fault. The Semantic Web for Earth and Environmental Terminology Ontology (SWEET), is the "de facto" broad-scope ontology for Earth and Environmental Sciences (EES) developed by NASA. It features incorporated unified fault entry description (DiGiuseppe et al., 2014). Conversely, the Commission for the Application and Management of Geoscience Information (CGI, 2006) group describes a fault observation as a fault type, movement type, and movement sense. In the case of the fault example, both sources may be needed and there may also be the necessity to extend or modify the source files to best fit the needs of the custom ontology. Building the ontology can cause decision-based issues within the workflow, as there might be several ways to present the data.

The Commission for the Application and Management of Geoscience Information (CGI) definition requires three files to cover the wide array of fault types and movements (Table 2). For this instance, the CGI ontology is more exhaustive than the SWEET ontology, yet the SWEET ontology covers a broader list of terminology for field sciences in general. Both may be used in conjunction, separately, or expanded upon with custom entries and represent an upper level of ontological information regarding geologic faults. Upper-level ontologies (i.e., domain-

independent ontologies) describing general concepts are proving useful within the biomedical, biological, and environmental science fields. Hybrid ontologies present a mix of foundational/core concepts and domain-specific knowledge bases.

Table 2. Comparison of two methods for describing a geologic fault.

Geologic feature type	Author	IRI*
Phenomena Geologic Fault	ESIP SWEET Ontology	http://cor.esipfed.org/ont?iri=http://sweetontology.net/phenGeolFault
Fault Type	CGI Geoscience Terminology	http://cor.esipfed.org/ont?iri=http://resource.geosciml.org/classifierscheme/cgi/2016.01/faulttype
Fault Movement Type	CGI Geoscience Terminology	http://cor.esipfed.org/ont?iri=http://resource.geosciml.org/classifierscheme/cgi/2016.01/faultmovementtype
Fault Movement Sense	CGI Geoscience Terminology	http://cor.esipfed.org/ont?iri=http://resource.geosciml.org/classifierscheme/cgi/2016.01/faultmovementsense

* Examples were obtained from the Community Ontology Repository (COR) (http://cor.esipfed.org/ont/#/). Accessed on April 2022.

Geologic relevant ontologies are available, such as the Environment Ontology (ENVO) that contains descriptions for environmental studies involving realms within ecosystems, processes, and scientific data qualities (Whetzel et al. 2016). Another resource is the Extensible Observation Ontology (OBOE), which captures semantics focused on scientific observations and measurements (Madin et al., 2007). The Semantic Sensor Network Ontology (SSN) describes sensors, actuators, observations for various scientific devices (Krötzsch et al., 2012). The Friend

of a Friend Vocabulary (FOAF, n.d.) ontology provides relationships for personal and working associations. The Ontology of Units of Measure (Golodoniuc, 2018) list standardizations for unit measurements used in scientific research. Regardless of the source(s) used, an exploration into the nature of the file is required to grasp the logical conventions used, each of which may result in a complicated effort by the user in terms of time, expertise, format, and analysis. Recent research by Zhan et al. (2021) reused select upper-level ontologies and described their interaction with proposed, lower-level ontologies supporting interpretation of historical geologic events inferred from field observations. This research aims at a similar approach of hybrid ontologies, selectively incorporating upper-level ontologies to facilitate use within geologic study sites.
In addition to observational concepts, an ontology supporting field geology needs to capture interpretative concepts.

Compton (1984) generalizes foundational information for a typical field notebook for any selected field site as containing data that is both evidence-driven and interpretative. This would also be true for an ontology intended to support geologic field mapping. Lisle & Barnes (1983) state that "… Fact must always be clearly distinguishable from inference." This key value is also displayed within Protégé, as the use of the HermiT reasoner (Glimm, 2014) clearly delineates machine-derived inferences with a yellow highlighted record bound by a dashed border since it is from a logical perspective. Searches may be performed strictly upon the explicit ontological data or may include the inferenced data, based on user's needs. If any of the inferences have been verified as fact, they can then be incorporated back into the ontology. Protégé provides this functionally via a simple button click per inference case.

Methods

Overview of Methods

This investigation used a case-based approach of a specific geology field site, Mt. Cristo Rey, in Doña Ana County, New Mexico. In collaboration with a field geology expert, two major competency questions were formulated that are fundamental to understanding a new field site that could be addressed by the system:

1. What geologic formation am I observing?
2. What geologic formation might I expect next?

Existing information relevant to the Mt. Cristo Rey study site was used to guide generation of a conceptual model of data utilized in the field for a variety of purposes. The conceptual model was implemented into a machine-readable ontology with encoded data. The ontology contains provenance information, typical field notes and map data, imported weather and earthquake data, and finally, expert domain data sourced from published journal articles describing geologic formations and stratigraphy for the area. The collaborating expert field geologist provided sample rock descriptions along with the correct identification of the geologic formation from which the sample was taken. Queries were constructed based on the rock descriptions to test the accuracy of the results generated by the system relevant to the competency questions.

Field Site

Mt. Cristo Rey is also known as “Cerro de Los Muleros” or “Cerro de Cristo Rey.” It is an Eocene andesite laccolith peak located within the southern Rio Grande rift valley just west of El Paso, TX (Lucas et al., 2010). This unit intruded a sequence of Cretaceous sedimentary rocks,

resulting in moderate to severe folding and faulting of these older sedimentary units. The oldest exposed sedimentary units are late Albian marine and non-marine sedimentary rocks that were deposited during a period of transgressive and regressive events. These are overlain by late Cretaceous rocks formed during development of the Chihuahua trough and opening the Gulf of Mexico (Hook, 2008; Lucas, 2010; Lovejoy, 1976). The fossil assemblage found within the shallow marine deposits provides a means to confidently identify different rock types and units suitable for geologic formation identification within the lithostratigraphy and sedimentary petrography.

KNOWLEDGE BASE DEVELOPMENT

A scope for the knowledge base was established based upon the two competency questions. In this case, the system was intended to assist with identification of a geologic formation from a description of its rocks and to use the stratigraphic column to provide end users with what formation should be expected next based upon what formation is being observed. The knowledge base developer should note the challenges associated with an increase in sensor data, user data, and modeling complexity, can represent an increase in biases, error, and uncertainty surrounding decision making. (Klein, 2015).

DATA SOURCES

Domain specific data for Mt. Cristo Rey was extracted from three peer-reviewed journal articles (Lovejoy, 1976; Belle, 1987; Lucas et al., 2010) using the online AI tool, ChatGPT (Adiwardana et al., 2020). These articles provide information of geologic formations in the area, the stratigraphic column, and depositional environments. ChatGPT generated brief summaries of relevant information of the given sections of the article, then was asked to create a master

summary based on all three article summaries. The last step was for ChatGPT to reduce the master summary into a keyword list. From this list of keywords, property assertions were made for each geologic formation in Protégé. This approach generated a non-biased domain expert dataset for competency testing purposes, without infringing upon intellectual property rights of the authors. All three sources are cited within the knowledge base. For example, for the major geologic features and properties of the geologic unit "Mancos" formation, ChatGPT was given the prompt to generate note the major geologic features from a selected text (i.e., the first text excerpt for each cited manuscript), then it was asked to add the geologic features found on each text and to combine all the features into a master list. The summarized results were in paragraph form, often with sections of directly quoted text without the citation to the author. ChatGPT was then prompted to generate a keyword list of the summary and the resulting list was comprised of single word features such as "shale" and "bivalves" or was shorten associations such as "Boquillas" as another name for the "Mancos" formation.

High-level weather, earthquake, and geological structure/processes data were sourced from existing ontologies. CGI and the SWEET ontologies provided geoscience terminology and semantic descriptions for the Earth Realm. NOAA's ontology was used for weather data and the USGS's QuakeML provided seismic data.

KNOWLEDGE BASE STRUCTURING

The class hierarchy was constructed using four major classes: 1) Field Data information meant to store in-situ field mapping information, such as recorded weather, earthquake, and local map observations and measurements; 2) Geologic information regarding geologic features such as time scale, stratigraphy, lithology, and depositional environments; 3) Imported ontological

data; and 4) Personal information about the authors/data collectors for purposes of maintaining provenance tracking. Subclasses within the major classes were subsumed into classes that created a hierarchy for stratigraphy, observations and measurements, and general geologic features. Instances, such as specific geologic formation names, were created within the class structure. The instances were linked to specific data relationships to further define concepts, for example associating formations to geologic structures and processes.

Using the "Add to ontology" function within the query box, a class within the main "Queries" can encode as "Class" features, "Object/Data Properties", "Instances" and "Annotations" were identified within the domain journal articles. Processing a section of domain expert data such as descriptions by Lucas et al. (2010) involves manually encoding of geologic concepts as ontology structure within the knowledge base. For example, a section of text that describes sandstone intervals each 0.4 m thick that contains Texigryhaea and displays ripple lamination exist in the upper part of the formation; would contain, as a concept, the class structures for formation stratigraphy, data property assertions for the thickness in meters, and individual/instances for "Texigryhaea" since the fossil name was not found in the imported ontologies, and "sandstone" which does exist within the CGI ontology is linked, and object property assertions for what ripple lamination "represents" (geologically). Annotation properties might be added to also note the general location within the unit these features are in (i.e., upper part). This process was derived uniquely by the ontology developer's interpretation of what was important and how to explicitly enter it into the ontology. Pre-existing ontologies imported into Protégé have their own structure. These were modified and manually linked to custom terms or preferred terms within the KB as needed.

Publishing of ontologies/knowledge bases requires the developer to follow copyright rules of expert domain data and take into consideration specific permission of use (e.g., licenses) based upon data sources used. The structure of ontology can be reused, and logical connections provided by property/data assertions by removing individual instances, which would then be populated by new knowledge base engineers for the specific data relating to the user's needs. This permits use of the ontology without including published content for a specific field site that might infringe upon intellectual property. Figure 3 shows custom instances linked to domain expert keywords for expanding instances to existing data included with in the imported ontology. For example, a siltstone-sandstone instance was created by adding a direct instance into the imported dataset and mapping the two individual instances as the "Same Individuals as" description for siltstone and sandstone, which were available from the imported dataset. Similar annotations can be made for instances as well as for property assertions. For example, Mesilla Valley (Member B) has an olive shale unit that can be expressed as an object property assertion annotation, allowing a more specific annotation.

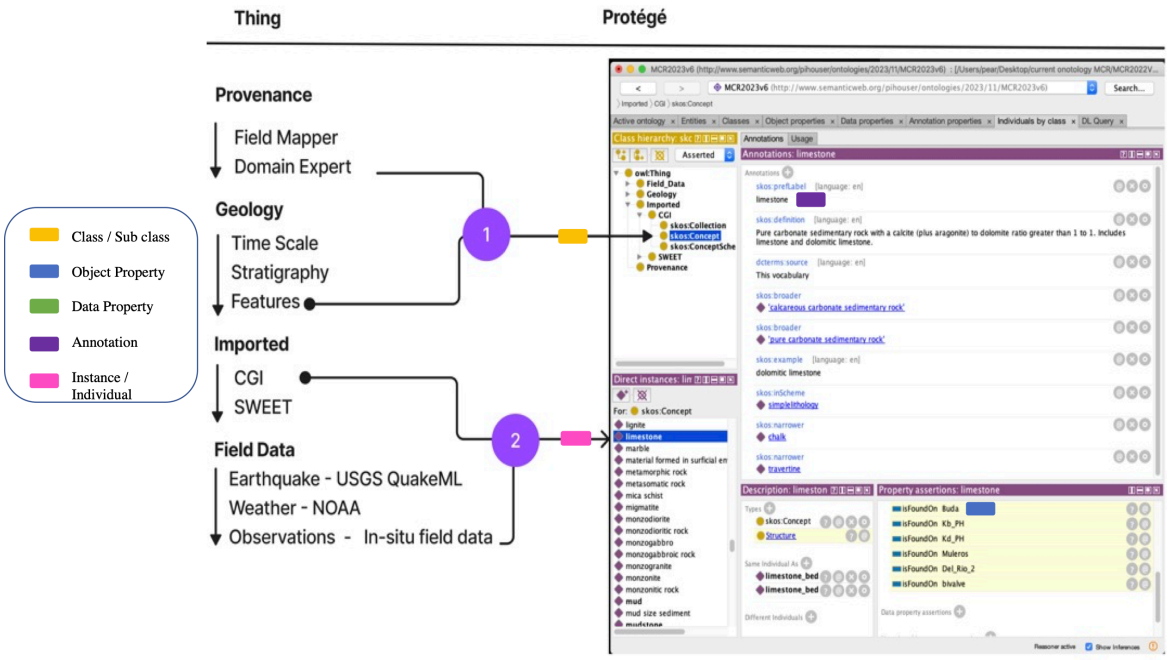

Figure 3: Example of integrating CGI class structure to custom entries.
Domain expert data from CGI (skos:lithology) containing annotations for a geologic feature are linked to custom entries of a geologic observation and measurements within the field data class.

Using the "Add to ontology" to specify explicit instance associations to other subclasses, for example linking to imported ontologies, enables the properties of the class, object, and data structures therein. Since there are imported source tags that are labeled differently than how the custom term (which is the preferred term), they are both linked to each other in the Annotation properties under the same range and domain and are sub-classes to the main preferred class name. The keyword: "source" is easier to read and use than the imported Dublin Core Terms "dcterms:source"; and "hasSource" is also added as a sub-class for easier typed queries (personal preference). However, use of the already included metadata standards for describing various aspects of a resource promotes interoperability and consistency across different systems. Each formation was manually assigned an integer number rank to represent its position within the stratigraphic column. This enabled the search results for "what unit might you expect next"

based on which unit is currently being observed. Additional links can expand upon other aspects of a formation, for example, the Mesilla Valley object property link to the “shale” class feature within the geosciml (CGI) imported ontology asserts links to annotations with pre-defined descriptions, definitions, and links to other concepts within CGI; in where “shale” is then automatically related to the CGI hierarchy of the class “mudstone” and its properties. Constraints and role chains are added to further define and allow for more complexity within the reasoner’s capabilities.

AUTOMATED REASONING BY THE KNOWLEDGE BASE

Logical considerations (in-terms of the Reasoner) are important when creating secondary relations within the ontology to make use of, such as role-chains and/or constraints are made during the data collection and integration process. Data relationships assign numerical values to objects and constraints force those numbers within a specific range given, for example strike and dip values are constrained to be integer values and fall within 0 – 360. Role chains provide a method of stacking logical relationships for an instance, chaining together the spatial relationships of the orders in stratigraphy for the reasoner to answer what unit might be expected next based on what unit is being observed.

The functional, inverse functional, transitive, symmetric, asymmetric, reflexive, and irreflexive characteristics determine inferences made by the reasoner. These characteristics help define relationships between instances within classes (and sub-classes) which require manual validation against the domain expert data if expert data verification is possible.

Logical checks within the ontology must be made by the user for nearly each item to verify that important information (or even possibly important information) is represented in the

ontology explicitly (via properties given or simply entered as an annotation) or implicitly via the reasoner. Logical connections made by the reasoner from improper characteristics selected upon creation dramatically increased the processing time for the reasoner and were corrected to maintain functionality as an "in-field" mapping tool. Reasoner (HermiT) Performance at 5,190 Axioms: 12 seconds (~12.34 secs) for the final ontology representation (see Table 3).

Table 3. Reasoner (HermiT) Performance: Axiom Count, Time, and Ontology development.

Axiom count	Time*	Description
5394 Axioms	~ 12.94 seconds	1 - 3 formations
5457 Axioms	~ 83.91 seconds	4 formations descriptions, property assertion error.
5546 Axioms	~ 252.8 seconds	5 formations descriptions, property assertion errors.
5555 Axioms	~ 135 seconds	6 formations descriptions, property assertion errors.
5568 Axioms	~ 8.92 seconds	7 formations descriptions, property assertion errors fixed.
5190 Axioms	~ 12.3 seconds	8 - 10 formations, Final ontology representation.

* HermiT Reasoner running on a 2012 Apple MacBook Pro, 2.5 Ghz Dual-core i5, 16 GB 1600 Mhz DDR3 RAM running macOS Catalina Version 10.15.17.

ONTOLOGY VERIFICATION AND USAGE

Protégé provides multiple levels of data verification both for syntax and logical errors. Access to the front-end graphical interface aids with no direct code development, and running the reasoner requires the ontology to be logically consistent within the relationships computed. Once the reasoner has initialized, verification of data is made by overviewing inferences and completing the competency questions backed by the domain expert data. Key terms and metadata from all the data sources (journal articles, field notes, existing ontologies, and reasoner logic) were reviewed using Noy's (2001) methodology to assess the structure, relationships, and

constraints appropriate for a lower-level ontology. Additional modification of the domain expert data is also incorporated into the ontology development to capture expected query results from the competency questions. For example, domain expert data strictly refers to the “shale” class feature within the Mesilla Valley geologic unit as “gray” within the object color property annotation, then search results will not trigger when using the non-domain expert statements that the color is “grey”; the required solution is to add a “sameAs” property links between the two colors (an example of a Functional property). Another example is linking the Anapra Sandstone formation (mentioned by one domain expert source) to the newer naming convention for the formation: the Mojado Formation. Mojado “hasThickness” is a reflexive property, as every formation has a thickness, and the thickness is set by the domain expert measurement or measurements recorded in the field. These properties assist the reasoner with proper and consistent inferences and answering DL queries.

Analysis

For this research, DL queries were used to validate the ontology content via the competency questions were and results were recorded separately in a spreadsheet. The searches include annotations within the logical axioms. The number of search hits of keywords from rock descriptions were tallied for each geologic formation found by the system. Since simple search results were given in real-time in response to typing, results were immediately assessed for accuracy and potential sources of error were considered.

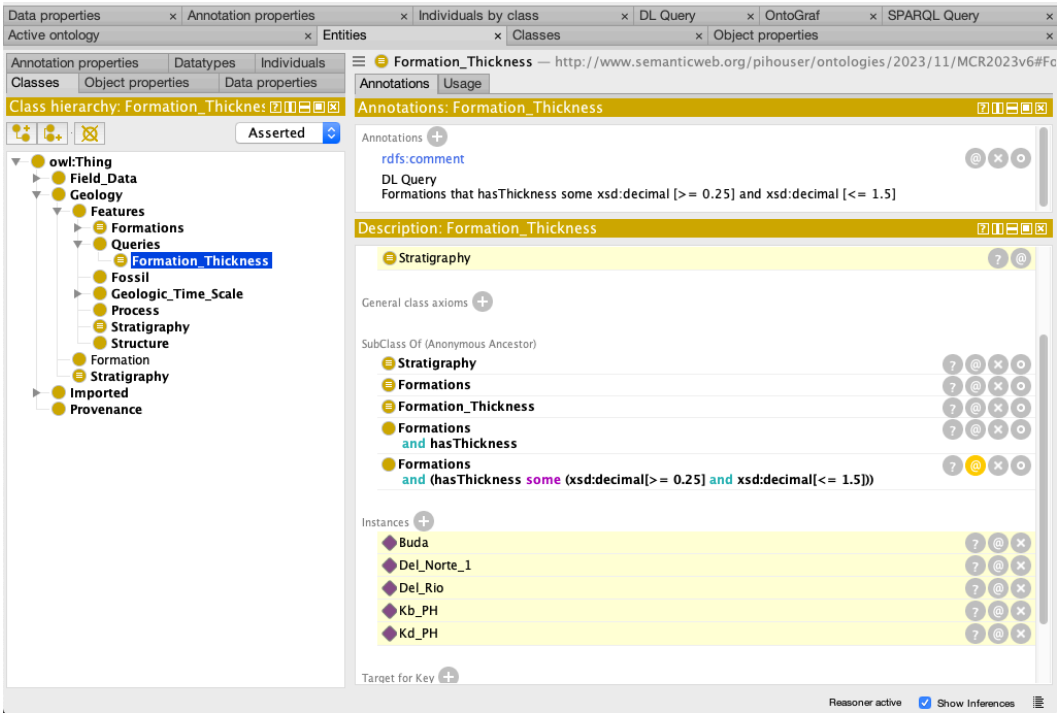

Figure 4: DL query example that can generate a class structure based on a query. Using the “Add to ontology” function within the query box, a class within the main “Queries” can encode as a “Formation_Thickness” sub-class where reasoner generated instances that meet the query requirements are automatically updated without having to write or know Dl Query syntax.

Figure 4 shows an example of a DL query statement that is encoded back into the ontology as the class "Formation Thickness” with reasoner-inferred instances, this makes for more efficient queries and requires less knowledge of the syntax requirements of DL query

statements. To assess the first competency question, “What geologic formation am I observing”, the field geology expert generated eleven rock descriptions for nine different rock units at the field site (Figure 5). These were illustrative of the kinds of descriptions we might expect to receive from students in a field geology class or someone who has a limited experience at a particular field site and/or may have limited access to the field site and will be collectively referred to as “end user descriptions” throughout the remainder of this thesis. The field expert provided the correct identification of the geologic formation being described. The descriptions were used to generate the key search terms queried by the system. The system returned a list of geologic formations that matched one or more of the search terms, with a count of the number of terms matched for each formation. The geologic formation with the highest number of matches was identified as the system result. This result was compared with the formation identified by the field geology expert and flagged as correct or incorrect. Correct and incorrect results were counted and calculated as a percentage of total results. Incorrect results were analyzed to identify issues encountered and possible solutions. To evaluate the second competency question of “Which geologic formation should I expect next?”, correct results from the first question were further analyzed for three selected cases. This was done with a mix of in-field verification and remote verification using the expert geologic map data of the study area (Figure 5).

The ontological use of the “expectNext” property relationship within the “ImObserving” property provided two possible inferred units – one stratigraphically above the current unit and one below. The actual next contact with another rock unit was visible by color change on the ground, which was then identified using the ontology. If that identification matched one of the two possible suggestions, a “correct” unit was selected and verified again with the expert map data. If the identified unit did not match either of the two possible inferred units, it would

represent a “red flag” for the end user that there must be a disconformity (or a non-conformity) that must be identified and mapped. An incorrect result for the first competency question necessarily resulted in incorrect results for the second competency question.

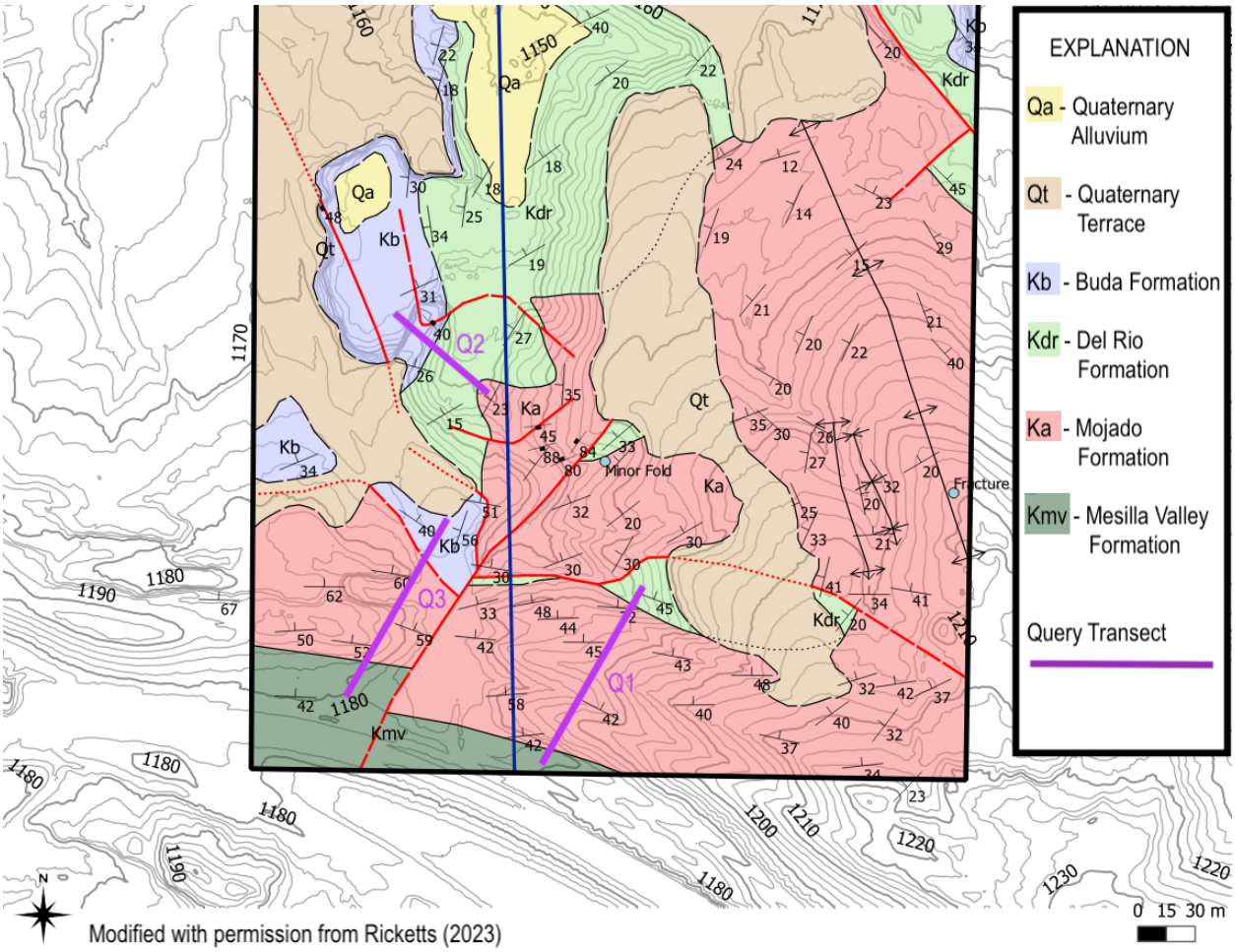

Figure 5: Geologic map provided by the field geology expert.
Lines in purple indicate where validation transects were taken (labelled Q1 & Q2), the midline is the unit being observed. Q3 provides an example where a fault has changed which unit might be expected next.

Results & Discussion

The ontology was able to correctly identify the correct geologic formation for 10 of the 11 end user rock descriptions. The last description (number 11) was a single word "shale" and since each unit contained shale, the count total resulted in all units being identified. Descriptions that provided more detail provided higher search results. The two DL Queries, which involved a more complex query statement were merged into the overall search results. A DL Query alone identified multiple geologic formations as the result, which alone would not be helpful in unit identification. However, when combined with the search results from the basic query, the DL Query did improve overall tallied results for the correct unit being identified. Results for all eleven rock descriptions are given in Table 4.

Table 4. Knowledge base tallies for combined keyword and DL Query searches for eleven end user rock descriptions within nine described rock formations.
The formation with the highest number of hits from the system was a correct answer if it matched the formation identified by the expert. MOJ=Mojado; DR=Del Rio; MES=Mesilla; MUL=Mulero; DN=Del Norte; BUD=Buda; FIN=Finlay; SM=Smeltertown.

EXPERT ID	MOJ	DR	MAN	MES	MUL	DN	BUD	FIN	SM
Mojado	13	6	2	8	3	4	1	0	2
Del Rio	10	12	6	6	8	10	9	11	5
Finlay	1	2	1	1	3	2	3	4	1
Buda	5	7	3	5	6	7	8	7	4
Del Rio	10	11	4	9	9	8	9	7	5
Mojado	6	3	2	3	3	3	2	2	2
Mesilla	9	6	4	11	6	7	4	3	6
Del Norte	5	6	3	5	5	7	5	3	4

Muleros	4	3	3	4	5	4	2	2	3
Smeltertown	6	6	3	6	6	6	4	2	7
Mancos	1	1	1	1	1	1	1	1	1

EXAMPLES

The first rock description was: "Light brown to orange medium-grained, well-sorted, well-rounded sandstone. Bed thicknesses range from 25 cm to 1.5 m and contain ripple laminations and trough cross-beds." Fifteen keywords were extracted from the description and used to search the knowledge base, including, for example, "light brown," "orange," "medium-grained," "well-sorted," "well-rounded," "sandstone," "ripple laminations," and "cross-beds." This search resulted in the correct domain expert answer: Mojado Formation with the highest tallied result (twelve). A DL Query was used to specify the bed thickness range, resulting in an additional three returns, including one for Mojado, raising its total to thirteen. The next highest tally was Mesilla Valley formation at 8 hits.

Searches involved subjective descriptions that involved alternative searches. Descriptions of rock color gave positive hits and negative (or not found) hits for "Orange." Color descriptions could potentially be described very differently from person to person resulting in varying success within search hits. Searches for hyphenated words, "well-sorted" did not find any matches, however a check on "well sorted" does correctly provide Mojado Formation as a correct hit. The same issue occurred with "medium-grained," requiring user-based logical checks that the system would otherwise miss. Another search issue arose for "well-sorted." A search for "well" and separately, "sorted", resulted in direct hits for incorrect formations that involved the inverse "poorly sorted", which logically could support the statement that non-listed formations that were

not labeled as “poorly sorted” could be inferred as “well sorted” or simple as not recorded/unknown. These searches were based off eliminations of units when the search for “sorted”, gave units that were poorly sorted, hence could not be the correct unit. This did give false positives for some of the other units along with the supported answer; however, these were not enough to impact the highest count totaled at the end. Ultimately, these were recorded as positive hits for the Mojado Formation, and the formation was still supported as the highest tallied return even with these inferences removed completely from the tally. Furthermore, removal of any alternate searches still correctly tallies Mojado as the highest search result.
DL Query also generated unexpected results. These results may not always give the correct result or provide a single supported result. Validating results require logical checks as well, as data may be correctly verified by the program may differ from correctly verified data from a scientific perspective. The description “25 cm to 1.5m” was units converted to meters. A search for bed range from 0 meters thick to a maximum of 2m resulted in no results; however, if the description had stated a maximum of 3.5m, the search results would have found the correct answer (Mojado), this represents a limitation within the dataset and not of the ontology or query mechanics.

A second example is the rock description: “Alternating sequence of light greenish yellow shale and greyish white nodular limestone beds. Limestone beds form ledges, are discontinuous, and range from 10-50 cm in thickness. Shale intervals contain isolated limestone nodules that coalesce up-section into more continuous beds. Limestone beds contain Exogyra fossils”. The domain expert answer is the Del Rio Formation. The knowledge base correctly identified the Del Rio formation. The keyword search alone did yield the correct result with eleven hits. The DL Query on bed thickness range also correctly identified Del Rio raising its tally of hits to twelve

but also identified three other formations: Del Norte and Buda, and Finlay. Finlay had the next highest tally of hits (eleven). Although the close counts between the correct rock unit (twelve) and closest rival (eleven) make it likely that the knowledge base could identify the wrong rock unit in other cases, the second competency question lends support to the Del Rio being correct in this example.

The second competency question, "what unit might I expect next?" was tested along three transects (Figure 7). Transects Q1 and Q2 represent locations with normal stratigraphic relationships that should be correctly identified by the knowledge base. Transect Q3 represents a location where a fault has disrupted the normal stratigraphic column such that an unexpected rock unit occurs. The knowledge base would be expected to correctly identify two possible rock units that could be encountered next, the one above and the one below the current rock unit. When the next rock unit is encountered in the field and it is identified using competency question 1, the mismatch between the expected and actual rock unit encountered would be a flag. The knowledge base performed correctly at all three transects.

An example is the Del Rio formation along Q2 (Figure 1). The Del Rio is stratigraphically bounded by the Buda Formation above and the Mojado Formation below. The Inference from: the SuperProperty of (role-chain), "imObserving (Del Rio) isBelow SubPropertyOf: expectAbove" correctly returns the Buda formation above and expects the Mojado formation next if going down stratigraphy. Observations in the field were made perpendicular to the strike of the unit being observed (Del Rio). This direction (southeast) enhances the likelihood of coming across a different geologic unit. Ground truth data confirmed that the unit being observed was Del Rio and walking to the contact to the southeast the knowledge base correctly identified that the next unit was the Mojado formation.

As mentioned above, the results for competency question Q2 can support the results from competency question Q1. In the second rock description above where the system selected Del Rio, but Finlay was a close second, traversing to the next rock unit as in Q2 and identifying that unit as Buda using the knowledge base would confirm that the correct rock unit was Del Rio. Finlay would not be present unless there were an unconformity present that would force the end user to explain the break in stratigraphy, which was not present at the field site locality (and thus not an issue for this example).

SECONDARY COMPETENCY QUESTIONS

The primary two competency questions were supplemented by testing several secondary competency questions related to the primary two. Since the knowledge base is consistently represented, the system can offer flexibility to ask more specific questions. The end user rock descriptions contained data that could be used to infer the following individual questions:

1. Which unit(s) contain fossils?
2. Which units have [some geologic feature]?
3. Which units have some thickness range?
4. Which age/period does the current unit belong to?
5. Are there any recorded field observations/measurements for a specific unit?
6. What features (such as color) belong to a unit?
7. What are a unit/observation's locality data? (e.g., Lat/Long/Elevation)
8. Who has mapped/recorded information about a certain unit?
9. What geologic processes relates to the geologic structural being observed?

Results from the structure and processes competency questions (1 through 8) were straight forward. The knowledge base contains explicit information linking a formation to a geologic feature. The process question (9) is derived by the reasoner by linking instances of geologic features with object property assertions with instances for the geologic processes. The process of deriving the answer to the question through the reasoner is preferred; it reduces work for setting multiple instances of a geologic process or a feature to a geologic unit by automatically creating those relationships with the reasoner. The processing time required for the reasoner to compute all the relationships was negligible, taking only a few seconds, which falls within the acceptable time frame for field work expectations. Links that extend beyond the expert domain data can be helpful during searches and/or queries. Object relationships such as, "sameAs" can bridge terminologies across domains or levels of expertise. For example, "volcanic rocks" is a layman term for "Igneous rocks." Having these linked in the knowledge base reduces the need to search for both terms and improves search results by providing both results. Linking structure to process means that when an instance of a geologic formation contains a structure the process will automatically be included, reducing the time required to explicitly enter these relationships.

CHALLENGES AND LESSONS LEARNED

While it is not necessary to have different people or groups to produce an ontology it is difficult for a single person to fulfill the roles of learning and maintain each of these tasks, however, it can still be done. It is preferable to have one person responsible for ontology creation, another for domain expert data, and someone else for the knowledge base development.

Integrating knowledge bases into geologic field mapping introduces possible steep learning curve. Familiarity of ontologies and reasoners, integration of data, and maintenance would be requirements to that workflow. Data capture and discovery results would also be incorporated into and from existing systems in use, such as mapping programs like QGIS and hand-written notes. Commitment to adopting XML (W3C, 2008), RDF (W3C, 2014), and OWL for common use as a digital library for cataloging field study sites would be beneficial for future use by subsequent users. Assimilation of these technological applications provides encoding for AI integration and the data consistency assist with interdisciplinary scientific studies. Short classes focused on knowledge base integration in field mapping, along with introduction into sematic web technologies, insight into semantic online repositories and integration into existing programs would also be beneficial with commitment of use with academic domains.

Validating possible logical decisions that are derived by the reasoner is important when establishing various aspects of class structure, instances, data or object properties. The interpretation of what is important and how to explicitly enter concepts into the ontology is a uniquely subjective process influenced by the developer's biases will determine the scope and impact of the knowledge base. Mesilla Valley Formation, for example, has six explicit relations entered for the instance (4 object relations, and 2 data relations). The rest were entered into other instances, which gave 28 inferred properties via the reasoner. Each formation and member were checked manually to ensure correct inferences were given. These logical checks go beyond the technical perspective to include possible logical errors within the scientific perspective.

False inferences were found due to incorrectly selecting an object property’s characteristics which relate instances within the class structure. This issue caused extra time wasted when running the reasoner, as hundreds of relationships were made incorrectly.

Adding extra information pertaining to rock unit/formation such as depositional environment and sedimentology can be challenging since there is more than one way to do this. Creating sub-classes for each type of information means that relational links to the proper formation are necessary, as information must be linked to the proper rock unit. Breaking down information can be both done via comment sections with the annotation section and/or listed individually as object/data properties. Since there are key terms for each of the sections (depositional and sedimentology), this presents more complications, as the user must decide whether to create new custom entries or search for and import existing ontologies that cover the required topics. Challenges involve creating instances which have varying degrees of specificity, such as linking outer-shelf environment that are only "rarely" impacted by storms. Here, an instance for "rarely impact" by storms, means that the link can be described as: 1) "Rarely affected by storms" as an instance; 2) with an annotation as "rarely impacted", or as 3) an object property of "rarelyImpactedBy" linking the feature to the formation. Considerations for query issues include formation thickness where unit thickness values may vary by observation by the data sources and exposed local geologic station where measured. Queries results based on inferences from the reasoner represent conditions where simple searches could not yield the same results. Saving the results and the query within the class hierarchy and as an instance would improve efficiency.

FUTURE RESEARCH DIRECTION

The thesis knowledge base offers several opportunities for future research. For instance, importing each geologic formation as a graph from the main ontology model would provide for efficient queries and flexibility for ontology reuse with other field site knowledge bases as a formation-specific ontology import. Another option is to integrate tasked based knowledge bases

for inferences into possible solutions for when unexpected rock units are found by relating to specific geological processes. This can be achieved by expanding upon the secondary competency questions of how a geologic structure is related to some specific geologic process, which can be based on property assertions such as “depositionalEnvironment”, “impactedBy”, and “represents”. Where a formation was “impacted” by some geologic process that “represents” some depositional environment that created the geologic structure being observed. These properties might be to link local geologic formations to other surrounding units to determine regional scale movements. Ultimately, the process for creating alternative versions for other field sites would still begin with Noy’s method for determining the competency questions to be answered.

Conclusions

This thesis investigated the use of semantic web techniques to support geoscience end users in the field. The thesis examines the potential of these techniques to generate inferences that can support in-situ geologic driven hypothesis development, a critical period where geologic observations and measurements can be uniquely verified while still within the study site.

To accomplish this goal, two competency questions were defined for ontology development: 1) identification of a geologic formation in the field based on descriptive characteristics of the rock units; and 2) given the observed geologic formation, and the known stratigraphic sequence, what formation should the end user expect next. A small sample section of a study site within Mt. Cristo Rey, New Mexico provided proof of concept for generation of a conceptual model of machine-readable field data utilized for a variety of common geologic driven questions. The thesis outlines the development of a knowledge base and the verification and validation of data sources to answer competency questions. The model implemented can be scaled up to handle larger field sites and/or integration of multiple field mapping data sources. Overall, the thesis aimed to analyze the potential for implementing knowledge-based techniques for supporting geoscience end users in the field that are not yet part of the common tools for field mapping. Results suggest that once developed a knowledge base can provide accurate support for end users in the field. However, development requires a major investment of time and a steep learning curve for both developers and end users. Providing a focused and constructive investment in teaching elements of cyberinfrastructure for use within geologic mapping could play a crucial role in achieving research goals commonly found within field studies. This investment would not only contribute to cutting-edge advancements in the field, but also would provide valuable skills for professionals working in geologic exploration and research.

References

Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedler, A., Thapa, K., ... & Kurach, K. (2020). Towards a human-like open-domain chatbot. Preprint arXiv:2001.09977

American National Standards Institute. (2021). Home. https://www.ansi.org/

Belle, E. R. (1987). A geochemical study of the organic matter within the lower cretaceous Mesilla Valley shale, Cerro de Cristo Rey uplift, Dona Ana County, New Mexico. *The University of Texas at El Paso*. Retrieved April 20, 2023, from https://www.proquest.com/docview/303629704?pq-origsite=gscholar&fromopenview=true

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 34-43.

Brown, J. S., & Adler, R. P. (2008). Minds on fire: Open education, the long tail, and learning 2.0. Educause review, 43(1), 16-32.

Christensen, G., & Knezek, G. (2019). Navigating the Fourth Industrial Revolution: Preparing Students for the Future. TechTrends, 63(6), 704-709. doi: 10.1007/s11528-019-00422-7

Commission for the Management and Application of Geoscience Information (CGI). (2006). About CGI. Retrieved May 1, 2023, from https://www.cgi-iugs.org/about-us/

Compton, R. R. (1985). Geology in the Field John Wiley & Sons. *New York*.

DiGiuseppe, N., Pouchard, L. C., & Noy, N. F. (2014). SWEET ontology coverage for earth system sciences. *Earth Science Informatics*, *7*, 249-264.

Domestic Names Committee of the U.S. Board on Geographic Names. (n.d.). U.S. Board on Geographic Names. Retrieved May 10, 2023, from https://www.usgs.gov/core-science-systems/ngp/board-on-geographic-names

Elkins, J. T., & Elkins, N. M. (2007). Teaching geology in the field: Significant geoscience concept gains in entirely field-based introductory geology courses. *Journal of geoscience education*, *55*(2), 126-132.

Foaf vocabulary. (n.d.). Retrieved April 6, 2023, from http://xmlns.com/foaf/spec/

Fonseca, F. T., Egenhofer, M. J., & Gilberto, C. (2002). *Using Ontologies for Integrated Geographic Information Systems*. *6*(3).

Glimm, B., Horrocks, I., & Motik, B. (2014). HermiT: An OWL 2 reasoner. Journal of Automated Reasoning, 53(3), 245-269.

Gil, Y., David, C. H., Demir, I., Essawy, B. T., Fulweiler, R. W., Goodall, J. L., & Yu, X. (2016). Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance. *Earth and Space Science*, *3*(10), 388-415.

Gil, Y., Pierce, S. A., Babaie, H., Banerjee, A., Borne, K., Bust, G., & Cheatham. M. (2018). Intelligent systems for geosciences: an essential research agenda. *Communications of the ACM,* 62(1), 76-84.

Golodoniuc, P., Tudorache, T., Nyulas, C. I., Noy, N. F., & Musen, M. A. (2018). The ontology of units of measure and related concepts. Journal of biomedical semantics, 9(1), 1-21. doi: 10.1186/s13326-018-0180-1

Hey, T., & Trefethen, A. E. (2005). Cyberinfrastructure for e-Science. *Science*, *308*(5723), 817 LP – 821. https://doi.org/10.1126/science.1110410.

Hook, S. C., (2008). Gallery of geology—Sierra de Cristo Rey: New Mexico Geology, 30, 93–94.

International Organization for Standardization. (2021). ISO - International Organization for Standardization. https://www.iso.org/home.html

Janssen, K., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. Information Systems Management, 29(4), 258-268. https://doi.org/10.1080/10580530.2012.716740

Kasenchak, R. T. (2019). What is Semantic Search? And why is it important? *Information Services & Use*, *39*(3), 205-213.

Klein, G. (2015). A naturalistic decision making perspective on studying intuitive decision making. *Journal of applied research in memory and cognition*, *4*(3), 164-168.

Krötzsch, M., Simancik, F., & Horrocks, I. (2012). A formal semantics for the Semantic Sensor Network ontology. Journal of Web Semantics, 15, 42-58. doi: 10.1016/j.websem.2012.02.003

Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., ... & Zhao, J. (2013). PROV-DM: The PROV data model. W3C recommendation, 30(4), 23.

Lisle, J. W. B. and R. J. (1983). Basic Geological Mapping. In *The Journal of Geology* (Vol. 91, Issue 2). https://doi.org/10.1086/628761

Lovejoy, E. M. P. (1976). Geology of Cristo Rey uplift, Chihuahua and New Mexico. *New Mexico Bureau of Geology and Mineral Resources Memoir*, *31*. Retrieved April 20, 2023, from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8

&ved=2ahUKEwjXsa6-x__9AhWMk2oFHWp2Ag4QFnoECBMQAQ&url=https%3A%2F%2Fgeoinfo.nmt.edu%2Fpublications%2Fmonographs%2Fmemoirs%2Fdownloads%2F31%2FMemoir-31.pdf&usg=AOvVaw0undptk6ktDdPA

Lucas, S. G., Krainer, K., & Spielmann, J. A. (2010). Cretaceous stratigraphy, paleontology, petrography, depositional environments, and cycle stratigraphy at Cerro de Cristo Rey, Dona Ana County, New Mexico. *New Mexico Geology*, *32*(4), 103–130.

Lundmark, A. M., Augland, L. E., & Jørgensen, S. V. (2020). Digital fieldwork with Fieldmove-how do digital tools influence geoscience students' learning experience in the field? *Journal of Geography in Higher Education*, *44*(3), 427-440.

Mitchell, M. (2019). Artificial Intelligence: A Guide for Thinking Humans. Farrar, Straus and Giroux.

Mookerjee, M., Vieira, D., & State, S. (2015). We need to talk: Facilitating communication between field- based geoscience and cyberinfrastructure communities. *GSA Today*, *25*(11). https://doi.org/10.1130/GSATG248GW.1

Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., & Villa, F. (2007). An ontology for describing and synthesizing ecological observation data. *Ecological informatics*, *2*(3), 279-296.

Musen, M. A. (2015). The Protégé project: A look back and a look forward. *AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence*, *1*(4). https://doi.org/10.1145/2557001.25757003

National Information Standards Organization. (2021). NISO - National Information Standards Organization. https://www.niso.org/

Neumann, K., & Kutis, M. (2006). Mobile GIS in Geologic Mapping Exercises. *Journal of Geoscience Education*, *54*(2), 153–157.

Noy, N., & McGuinness, D. L. (2001). Ontology development 101. *Knowledge Systems Laboratory, Stanford University*, *2001*.

Open Geospatial Consortium. (2021). OGC - Making Location Count. https://www.ogc.org/

Parsia, B., Sirin, E., & Cuenca, B. (2012). OWL 2 Web Ontology Language Manchester Syntax (Second Edition). W3C Recommendation. https://www.w3.org/TR/owl2-manchester-syntax/

Pennington, D., Ebert-Uphoff, I., Freed, N., Martin, J., & Pierce, S. A. (2020). Bridging sustainability science, earth science, and data science through interdisciplinary education. *Sustainability Science*, *15*, 647-661.

Peters, S. E., Husson, J. M., & Czaplewski, J. (2018). Macrostrat: A platform for geological data integration and deep-time earth crust research. *Geochemistry, Geophysics, Geosystems*, *19*(4), 1393-1409.

Plale, B., McDonald, R. H., Chandrasekar, K., Kouper, I., Konkiel, S., Hedstrom, M. L., & Kumar, P. (2013). SEAD virtual archive: Building a federation of institutional repositories for long-term data preservation in sustainability science.

Priestnall, G., & Polmear, G. (2007). A synchronized virtual environment for developing location-aware mobile applications. *Proceedings of the 15th Annual Conference on Advances in Mobile Computing and Multimedia. Retrieved on October 24, 2022, from* http://www.geos.ed.ac.uk/~gisteac/proceedingsonline/GISRUK2007/PDF/4A3.pdf

Saini-Eidukat, B., Schwert, D. P., & Slator, B. M. (2002). Geology explorer: Virtual geologic mapping and interpretation. *Computers and Geosciences*, *28*(10), 1167–1176. https://doi.org/10.1016/S0098-3004(02)00036-5

Scianna, A., & Ammoscato, A. (2010). 3D GIS DATA MODEL USING OPEN SOURCE SOFTWARE. *ISPRS Archive Vol. XXXVIII, Part 4-8-2-W9, "Core Spatial Databases - Updating, Maintenance and Services – from Theory to Practice", Haifa, Israel, 2010*, *XXXVIII*, 120–125.

Sinha, a. K., Malik, Z., Rezgui, A., Barnes, C. G., Lin, K., Heiken, G., Thomas, W. a., Gundersen, L. C., Raskin, R., Jackson, I., Fox, P., McGuinness, D., Seber, D., & Zimmerman, H. (2010). Geoinformatics: Transforming data to knowledge for geosciences. *GSA Today*, *20*(12), 4–10. https://doi.org/10.1130/GSATG85A.1

Smith, J., & Jones, A. (2021). The Manchester Syntax: A Practical Guide for Ontology Developers. Journal of Semantic Technology, 10(2), 45-57.

Smith, P., Malik, T., & Berg-Cross, G. (2016). Rediscovering EarthCube: Collaborate. Or collaborate not. There is no I. *Digital Library Perspectives*, *32*(3), 153–191. https://doi.org/10.1108/DLP-09-2015-0017

Swetanisha, S., Panda, A. R., & Behera, D. K. (2022). Land use/land cover classification using machine learning models. *International Journal of Electrical & Computer Engineering (2088-8708)*, *12*(2).

U.S. Geological Survey. (n.d.). Home page. Retrieved May 9, 2023, from https://www.usgs.gov/

Walker, J. D., Tikoff, B., Newman, J., Clark, R., Ash, J., Good, J., Bunse, E. G., Möller, A., Kahn, M., Williams, R. T., Michels, Z., Andrew, J. E., & Rufledt, C. (2019). StraboSpot

data system for structural geology. *Geosphere*, *15*(2), 533–547. https://doi.org/10.1130/GES02039.1

Whetzel, P.L, Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., & Musen, M. A. (2016). The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. Journal of Biomedical Semantics, 7(1), 57.

World Wide Web Consortium. (2008). Extensible Markup Language (XML) 1.0 (Fifth Edition). https://www.w3.org/TR/REC-xml/

World Wide Web Consortium. (2014). RDF 1.1 Concepts and Abstract Syntax (Second Edition). https://www.w3.org/TR/rdf11-concepts/

World Wide Web Consortium. (2023). About W3.org. Retrieved August 8, 2022, https://www.w3.org/Consortium/

Zhan, X., Lu, C., & Hu, G. (2021). Event sequence interpretation of structural geological models: A knowledge-based approach. *Earth Science Informatics*, *14*, 99-118.

Vita

Perry I. Houser worked as a Systems Engineer with T-Mobile Wireless in Albuquerque, NM before completing his associate degree in Computer Science from the El Paso Community College (EPCC), and then completing his Bachelor's in Science from The University of Texas at El Paso. He worked with the Ysleta del Sur Tigua Tribe on historical, travel and tourism exhibits, websites, and a book publication. He has given several talks for the Upwards Bound federal program for low-income and first-time entry into college students at EPCC. He focuses on outreach opportunities and has given talks to El Paso ABC 7 News about local geology sites, started a science and technology club at UTEP, along with participation in Earth Science Day and volunteer science fair judge. During his undergraduate degree he did field research in the Arctic and British Columbia. During his graduate work, he has worked as a research assistant and as a teacher assistant and has received certificates in a GIST program and the OSINT Pathfinder Program. To further his degree, he has been awarded funding through NASA research grants, geology department grants and tuition endowment grants that have helped with research and workshops with National Center for Atmospheric Research, Texas Advance Computer Center, and Earth and Environmental Research. He is on six co-authored abstract publications and presented in nine conference posters for organizations such as American Geophysical Union, Geologic Society of America, Incorporated Research Institutions for Seismology, Lunar and Planetary Science Conference, Year of the Solar System, and the UTEP Geological Sciences Colloquium, where he was also able to serve on the committee. He has given conference talks for ESRI User Conference and the National Security Studies Institute – ANS Conference. He worked at an internship with USC IS-GEO: Intelligent Systems with Dr. Yolanda Gil.

Contact Information: pihouser@miners.utep.edu