Ontology-Driven Discovery of Scientific Computational Entities

Pearl W. Brazier

University of Texas at El Paso, brazier@utpa.edu

Follow this and additional works at: https://digitalcommons.utep.edu/open_etd

Part of the Computer Sciences Commons

Recommended Citation


https://digitalcommons.utep.edu/open_etd/2445

This is brought to you for free and open access by DigitalCommons@UTEP. It has been accepted for inclusion in Open Access Theses & Dissertations by an authorized administrator of DigitalCommons@UTEP. For more information, please contact lweber@utep.edu.
ONTOMETRY-DRIVEN DISCOVERY OF SCIENTIFIC COMPUTATIONAL ENTITIES

PEARL W. BRAZIER

Department of Computer Science

APPROVED

Ann Q. Gates, Ph.D., Chair

Vladik Kreinovich, Ph. D.

Randy Keller, Ph.D.

Artem Chebotko, Ph.D.

Patricia D. Witherspoon, Ph.D.
Dean of the Graduate School
COPYRIGHT BY

PEARL W. BRAZIER

2010

All Rights Reserved
DEDICATION

To my husband Gerald (Jerry) Brazier, and my children, Michael Brazier, Jenny Turner, her husband Jonathon, and my two grandchildren, Avery and Grant Turner.
ONTOGONY-DRIVEN DISCOVERY OF SCIENTIFIC COMPUTATIONAL ENTITIES

By

PEARL W. BRAZIER

DISSERTATION

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

DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE
THE UNIVERSITY OF TEXAS AT EL PASO
AUGUST 2010
ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Ann Q. Gates for making it possible to pursue the degree in a non-traditional way. I am extremely grateful for her unselfish support and understanding over the years that is has taken me to finish my degree. I consider her more than my advisor but also a good personal and professional friend. I especially value our sharing of our experiences as faculty and women in computer science and I have great respect for her accomplishments. She is an inspiration to me. She has taught me so much about doing research. I also owe a tremendous debt of gratitude to Dr. Artem Chebotko, UTPA faculty serving on my committee, for believing in my research direction, guiding me to expand the direction, and spending countless hours sharing his research and expertise with me. It would not have been possible without him. I also thank my other committee members, Dr. Vladik Kreinovich, for his support and expertise and Dr. Randy Keller for sharing his geology experience. Thanks also go to other UTEP faculty for their friendship and guidance and their willingness to have me in their classes remotely. Special thanks go to Dr. Xiaohui (Sofie) Wang and Dr. Gerald Brazier, UTPA faculty, for their statistical expertise. Many people at the University of Texas – Pan American have supported me through the years, Dr. Edwin LeMaster, Dean of College of Science and Engineering, Dr. Zhixiang Chen, UTPA Chair of Computer Science Department, UTPA Computer Science faculty and staff, other UTPA faculty, and UTPA students. I especially want to thank UTPA graduate students, Liliana Ochoa for help in the wiki implementation of GEO-SEED and in the design of the GEO-SEED logo, Anthony Piazza, Eric Gonzalez, and Andrey Kashlev for their help with the performance study setup, and Maria Garza, who was involved in my
early research during her Computer Science Master’s under my supervision. There have been many friends who have encouraged me in this endeavor. And last, but most important, I thank my husband for his enduring love and support through the years as I have pursued my dreams. My children and grandchildren have been the joy of my life. I am so proud of their achievements. My extended family has been the source of making me who I am. In particular, I remember my mother for supporting me in so many ways.

I would like to acknowledge the support of the NSF Award EIA-0080940: “Graduate Education for Minority Students in Computer Science and Engineering: Extending the Pipeline”, a Minority Institutions Infrastructure (MI-I) grant, and NSF HRD-0734825.
ABSTRACT

Many geoscientists use modern computational resources, such as software applications, Web services, scientific workflows and datasets that are readily available on the Internet, to support their research and many common tasks. These resources are often shared via human contact and sometimes stored in data portals; however, they are not necessarily available for metadata annotations that can assist in collaborative research or machine processing. Scientists’ knowledge of their resources and processes are at risk of being lost. A scientist-driven discovery environment, which assumes a level of technical expertise generally possessed by a geoscientist, is needed to enable discovery and comparisons of computational entities. The goal of the research was to investigate an ontology-driven discovery approach that can be distributed on the Web and that can support the elicitation, documentation, and registration of computational entities and other resources. The main research efforts included: definition of a portal architecture that supports the registration, annotation, knowledge extraction and management, and discovery of computational resources using an ontology-driven approach; and evaluation of the usability and performance of a prototype system based on the architecture. The resulting innovative architecture blends Web 2.0 and Semantic Web technologies, features an intuitive and collaborative work environment of a structured wiki, and machine-interpretable metadata accessible via standard Semantic Web languages, such as RDF, SPARQL, and OWL. The ontology, called Computational Entity Discovery Ontology, provides a standard vocabulary for metadata acquisition encoded in standard Semantic Web languages OWL and RDF. The developed wiki is a next-generation structured wiki that not only delivers Web 2.0 advantages to provide a collaborative knowledge management environment, but also presents information in a well-arranged, structured fashion. The designed relational RDF repository features a unique user-customizable database schema generation approach based on the idea of horizontal partitioning. The experimental comparison with two known schema-oblivious and schema-aware database storage schemes showed superior performance of the proposed approach.
# TABLE OF CONTENTS

DEDICATION ........................................................................................................................................................ iii
ACKNOWLEDGEMENTS ........................................................................................................................................... v
ABSTRACT ............................................................................................................................................................ vii
LIST OF TABLES ................................................................................................................................................... x
LIST OF FIGURES ................................................................................................................................................ xi
CHAPTER 1: INTRODUCTION ........................................................................................................................ 1
1.1 Overview .................................................................................................................................................... 1
1.2 Motivation .................................................................................................................................................. 2
1.3 Research Objectives ............................................................................................................................... 3
1.4 Significance of the Contribution ....................................................................................................... 4
1.5 Organization of the Dissertation ...................................................................................................... 6
CHAPTER 2: BACKGROUND ........................................................................................................................... 7
2.1 Cyberinfrastructure ............................................................................................................................... 7
2.2 Relevant Cyberinfrastructure Support ............................................................................................... 10
2.3 Semantic Web Technologies ........................................................................................................... 13
2.4 Related Work ......................................................................................................................................... 17
CHAPTER 3: SCIENTIFIC COMPUTATIONAL ENTITY DISCOVERY ONTOLOGY .................... 21
3.1 Use Cases ................................................................................................................................................. 21
3.2 The Scientific Computational Entity Discovery Ontology Concept ........................................... 26
3.3 Relationships in the Ontology Concept ............................................................................................ 28
3.3 The Ontology Profile Concept ........................................................................................................ 29
3.4 Relationship Structure in the Ontology ......................................................................................... 38
CHAPTER 4: GEOSCIENCES WEB SERVICES METADATA MANAGEMENT AND DISCOVERY SYSTEM .......................................................................................................................... 39
4.1 GEO-SEED Architecture .................................................................................................................... 41
4.2 Structured Wiki Implementation ........................................................................................................ 45
4.3 GEO-SEED RDF Repository ............................................................................................................ 45
CHAPTER 5: EVALUATION AND ASSESSMENT ................................................................................... 52
5.1 GEO-SEED Usability Study-I ........................................................................................................... 52
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2</td>
<td>GEO-SEED Usability Study-II</td>
<td>55</td>
</tr>
<tr>
<td>5.3</td>
<td>Performance Study</td>
<td>62</td>
</tr>
<tr>
<td>5.4</td>
<td>Study Results</td>
<td>67</td>
</tr>
<tr>
<td>6.1</td>
<td>Conclusions</td>
<td>69</td>
</tr>
<tr>
<td>6.2</td>
<td>Future Work</td>
<td>70</td>
</tr>
<tr>
<td>REFERENCES</td>
<td></td>
<td>71</td>
</tr>
<tr>
<td>APPENDIX A</td>
<td>SCREEN SHOT OF COMPLETED GEO-SEED ENTRY</td>
<td>80</td>
</tr>
<tr>
<td>APPENDIX B</td>
<td>SURVEY OF GEOSCIENCE RESEARCHERS</td>
<td>83</td>
</tr>
<tr>
<td>APPENDIX C</td>
<td>GEO-SEED Usability Study I</td>
<td>85</td>
</tr>
<tr>
<td>APPENDIX D</td>
<td>USABILITY STUDY II</td>
<td>92</td>
</tr>
<tr>
<td>D.1</td>
<td>GEO-SEED Usability Study</td>
<td>92</td>
</tr>
<tr>
<td>D.2</td>
<td>Usability Survey Results</td>
<td>94</td>
</tr>
<tr>
<td>D.3</td>
<td>Descriptive Statistics for Usability Survey</td>
<td>102</td>
</tr>
<tr>
<td>APPENDIX E</td>
<td>PERFORMANCE STUDY</td>
<td>108</td>
</tr>
<tr>
<td>CURRICULUM VITAE</td>
<td></td>
<td>121</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1: Profiles of Scientific Computational Entities ................................................................. 31
Table 2: GeneralProfile Descriptors .............................................................................................. 33
Table 3: OoS Profile Descriptors ..................................................................................................... 34
Table 4: Invocation Profile Descriptors .......................................................................................... 35
Table 5: Deployment Profile Descriptors ....................................................................................... 36
Table 6: Implementation Profile Descriptors .................................................................................. 37
Table 7: Geoscience Profile Description ......................................................................................... 37
Table 8: Extracted RDF Triples .................................................................................................... 46
Table 9: Triple Table ....................................................................................................................... 47
Table 10: Property Table for type ................................................................................................ 48
Table 11: Property Table for trust .................................................................................................. 48
Table 12 Property Table for availability ......................................................................................... 48
Table 13: Profile Table for GeneralProfile .................................................................................... 49
Table 14: Profile Table for QoSProfile .......................................................................................... 49
Table 15: Time to Compute Queries .............................................................................................. 68
LIST OF FIGURES

Figure 1: Geoscience Use Case Diagram...................................................................................................... 22
Figure 2: Scientific Computational Entities Discovery Ontology Graph................................................. 28
Figure 3: GEO-SEED Scientific Computational Entity Discovery Ontology ........................................... 30
Figure 4: Instances of Computational Entities .......................................................................................... 38
Figure 5: Researcher’s Discovery Methods ................................................................................................. 40
Figure 6: Researcher’s Sharing Methods .................................................................................................... 40
Figure 7: GEO-SEED Architecture ............................................................................................................. 43
Figure 8: Gravity Contour Map Workflow .................................................................................................. 53
Figure 9: Current Knowledge of Web Related Technologies .............................................................. 57
Figure 10: Overall Rating of GEO-SEED Usability ................................................................................. 58
Figure 11: Ease of Using GEO-SEED to Register Computational Entities ............................................. 59
Figure 12: Degree of Difficulty Registering Information into Descriptor Fields .................................... 60
Figure 13: Ease of Ability to Use GEO-SEED Features ......................................................................... 61
Figure 14: Degree of Helpfulness of GEO-SEED ................................................................................... 62
Figure 15: RDF Repository Query Performance ....................................................................................... 68
Figure 16: Screen Shot Sample GEO-SEED Entry .................................................................................. 80
Figure 17: Screen Shot GEO-SEED Entry – General Information ...................................................... 81
Figure 18: Screen Shot GEO-SEED Entry - Invocation and Quality of Service Information .... 81
Figure 19: Screen Shot GEO-SEED Entry - Deployment, Implementation, and Domain Knowledge Information ............................................................................................................. 82
CHAPTER 1
INTRODUCTION

1.1 Overview

Many geoscientists use modern computational resources, such as software applications, Web services, scientific workflows and datasets that are readily available on the Internet, to support their research and many common tasks. These resources are often shared via human contact and sometimes stored in data portals; however, they are not necessarily available for metadata annotations that can assist in collaborative research or machine processing. Scientists' knowledge of their resources and processes are at risk of being lost. A scientist-driven discovery environment, which assumes a level of technical expertise generally possessed by a geoscientist, is needed to enable discovery and comparisons of computational entities, as well as to allow domain experts in science to easily convert existing information, knowledge, methods, and workflow integrated processes. There are numerous technologies that can support such an environment: an open set of standards and protocols supported by the World Wide Web Consortium (W3C) (W3C, 2009) and Semantic Web Services Framework (SWSF) (SWSF, 2005), which utilizes Web 2.0 (Social_network, 2009) and the workflow model of the Semantic Web and provides services to the scientific community. This dissertation is focused on geoscientists and their need to document scientific computational entities, such as Web services, scientific methods, software applications, and workflows, with the aim of facilitating the discovery and reuse of such entities. The goal of the research is to investigate an ontology-driven
discovery approach that can be distributed on the Web and that can support the elicitation, documentation, and registration of computational entities and other resources.

1.2 Motivation

The search and discovery capabilities of existing geosciences repositories such as GEON (GE008) and PACES (PAC) depend on metadata that describes available resources, its representation, expressiveness, storage and query interfaces. While a geoscientist can locate a required resource, she or he would have to manually register it with a Scientific Workflow Management System (SWFMS) (SWFMS, 2010) in order to use it in a scientific workflow design and execution. The integration of data and software repositories with SWFMS’s is an open problem that requires the exchange of machine-interpretable metadata acquired via a uniform description vocabulary. Currently, existing repositories are largely intended for humans rather than machines or automated agents. The current repositories also lack the capability of users to annotate the resources with ratings for the quality of service attributes, which would be of value to future users.

Geosciences researchers have numerous computational entities that support many common tasks. These processes are often shared via human contact and sometimes stored in data portals; however, they are not necessarily available for metadata annotations that can assist in collaborative research, or machine processing. Scientists’ knowledge of the processes they use is at risk of being lost. A scientist-driven discovery environment, which assumes a level of technical expertise generally possessed by a geoscientist, must be
developed to enable the user to discover and compare computational entities with support of ontologies if needed.

1.3 Research Objectives

The research objectives defined to meet the goal of defining an ontology-driven discovery approach of computational entities are described below.

Objective 1. Define an ontology for scientific computational entities that supports the development of a repository and a system that can retrieve computational entities.

*Activities:*

A. Define use cases for the ontology.

B. Determine the essential elements of an ontology that documents the features and relationships used to identify computational entities and distinguish one from another.

Objective 2. Define an architecture that supports an ontology-driven approach.

*Activities:*

A. Investigate the relationships of registration, annotation, and knowledge extraction.

B. Investigate efficient approaches for storing information.

Objective 3. Evaluate the usability of a system based on the ontology-driven discovery approach with respect to navigation, ease of registering as a user, ease of document ing computational entities, and ease of searching and retrieving computational entities.
Activities:

A. Design and implement a prototype system based on the ontology-driven discovery approach.

B. Conduct a usability study of the prototype system with computer scientists and geoscientists (novices and experts).

Objective 4. Evaluate the performance of a system based on the ontology-driven approach.

Activities:

A. Design the schema and implement a relational RDF repository that supports efficient storage and querying of documented scientific computational entities based on the ontology-driven discovery approach.

B. Run a simulation to analyze the performance of the system.

1.4 Significance of the Contribution

The work will enhance the capabilities of the researcher to collaborate, share and evaluate their processes, and potentially automate the workflow composition of their computational entities. The work contributes to the charge from the Semantic Web community to find application areas that makes use of Web 2.0 and the Semantic Web technologies in order to realize the potential of these technologies (Chen, 2004). The scientific community is experiencing an avalanche of data, which has driven the scientist to seek ways to manage and use the data in their research. This has been defined as “Big Science”, i.e., scientists working with massive computational power and volumes of data. Utility computing –
computational resources on demand - and virtual organizations will also impact the capabilities of the scientific community to not only use this mass amount of data, but also to use this data to form hypotheses for new research. This capability is now being made available in the realm of cloud computing and social networking (De Roure, 2010).

The scientific computational entity discovery ontology presents a design that will enable the scientist to share the computational entities they create and use. This led to the design of a prototype, which makes use of Web 2.0 technologies that can be distributed on the Web. This will support the scientist for their need to record and share their important computational entities. Scientists can also share the processes they use, which might be expressed as scientific workflows. These may be pre-existing entities, but the design also supports the sharing of new results. In addition, the design uses the widely accepted RDF storage for the information. The characteristics of the organization of the information needed to support the sharing and retrieval of their computational entities led to a novel design based on horizontal-partitioning for organizing the RDF data. This design uses SPARQL queries for the RDF representation using a SPARQL-to-SQL translation (Chebotko, 2009), allowing storage of the RDF data in a traditional RDBMS. New to the dissertation is the result that the novel design performs significantly better for commonly used queries to retrieve desired computational entities. A performance study comparison with two known schema-oblivious and schema-aware database storage schemes showed superior performance for the proposed user-customizable schema storage approach. The usability study of GEO-SEED supports the usability of the wiki design.
More recently, the Web Linking Open Data (LOD) Project (LOD, 2008) has been formed with the goal of moving the Web from the idea of separated documents to a wide information space of data. To accomplish this, Web resources must be easier to discover, more valuable, and easier for people to reuse. The research reported in this dissertation fits right into these arenas. The research also contributes to the efforts described in NSF’s National Cyberinfrastructure (CI) Vision for 21st Century Discovery (Rep09). One of which is to provide a sustainable CI that is secure, efficient, reliable, accessible, usable, and interoperable, and which evolves as an essential national infrastructure for conducting science and engineering research and education.

The framework will be useful for not only the geosciences domain, but is general enough to allow extension to other domains.

1.5 Organization of the Dissertation

The remaining chapters of the dissertation are organized as follows. Chapter 2 gives background information from e-science and scientific workflows, Scientific Workflow Management Systems, Web 2.0 Technologies, Semantic Web Technologies, and related works. Chapter 3 describes the Computational Entity Discovery Ontology. Chapter 4 presents the GEO-SEED pilot wiki and RDF Repository. Chapter 5 presents the evaluation and assessment of the efforts. And lastly, Chapter 6 presents the conclusions and future work.
CHAPTER 2
BACKGROUND

This chapter discusses background information needed to set the context of the research. To develop a scientist-driven Web-based registration and discovery system, Web 2.0 and Semantic Web technologies are well suited technologies. Another consideration frequently overlooked in the literature is the efficient storage of scientific information that is understandable by machines. In order to support reusability of computational entities, information must be stored using widely accepted standards, e.g., RDF. Because the overarching goal of the proposed approach is to contribute to the cyberinfrastructure in the geosciences, this chapter begins with a discussion of cyberinfrastructure, including a discussion of e-science and the state of cyberinfrastructure in the geosciences. The next section focuses on cyberinfrastructure support that is relevant to the information defined by the computational scientific entity ontology followed by a description of the Semantic Web technologies that are required to build the prototype for the ontology-driven discovery system. The prototype implemented for purposes of this research is GEO-SEED (GeoScienceWebServiceDiscovery). This chapter ends with related work.

2.1 Cyberinfrastructure

As technologies have evolved, capabilities now exist for scientists to take advantage of cyberinfrastructure. The capacity of this technology, computing systems, data, information resources, networking, digitally enabled-sensors, instruments, virtual organizations, and observatories, along with an interoperable suite of software services and tools, has crossed
thresholds that now make possible a comprehensive "cyberinfrastructure" on which to build new types of scientific and engineering knowledge environments and organizations and to pursue research in new ways and with increased efficacy (De Roure, 2009). In cyberinfrastructure, scientific data repositories, semantic mediation services, and scientific workflows have become key players in supporting modern in-silico experiments that can lead to important scientific discoveries (Lausen, 2005) (Chebotko, 2007) (De Roure, 2007).

The term e-Science has been used to describe computationally intensive science that is carried out in highly distributed network environments, or science that uses immense data sets that require grid computing (E-Science, 2007). This term has largely been used for research in the UK and Europe. United States - based initiatives typically use the term cyberinfrastructures (Cyberinfrastructure, 2010).

In geosciences, rich data repositories, such as, Geosciences Network GEON (GEO08) and the Pan-American Center for Earth and Environmental Studies (PACES) (PAC), are readily available to geologists. Using these collaborative portals, scientists are able to search for existing tools and datasets as well as register their own assets. However, researchers in the geosciences use processes to support many common tasks and these have been largely shared via human contact and are not necessarily available for easy sharing, in particular, in a form suitable for automated machine-enabled sharing.

Geoscientists have made strides in developing web portals to collaborate and share their data and tools. GEON is an open collaborative project that is developing a cyberinfrastructure for integration of three and four-dimensional earth science data (GEO08). This portal supports resources for Topography, Geophysics, and Geology.
Geologists have access to the GEONGrid portal, which enables them to register their data sets, tools, Webservices, Web mapping services, and ontologies, providing access control to classes of users. Registered users are provided with search capability, data backup, mapping service for shape files, and data integration service for relational databases. The portal supports the inclusion and viewing metadata. Each entry has fields to include a title, file format, author, spatial coverage, description, metadata, and usage statistics.

The Pan American Center for Earth and Environmental Studies (PACES) supports another portal targeted more specifically to the geologist, and is dedicated to the storage and processing of gravity data (PAC). The data at this portal is based on U.S. and the border region data sets that have been compiled from a vast number of sources. The database is part of a collaborative effort by the U.S. Geological Survey, University of Texas at El Paso, and Arizona State University. This portal primarily allows the scientist to register new datasets to make them available for other scientists to use, but does not support the registration of the range of capabilities of the GEON portal. The PACES data sets have been registered with GEON.

The search and discovery capabilities of the repositories depend on metadata that describes available resources, its representation, expressiveness, and storage and query interfaces. While a geoscientist can locate a required resource, to incorporate it into a scientific workflow, he or she has to manually register it with Scientific Workflow Management Systems. A more detailed discussion of Scientific Workflow Management Systems is included in the next subsection. The integration of data and software repositories of these systems is an open problem that requires the exchange of machine-
interpretable metadata acquired via a uniform description vocabulary. Currently, existing repositories are largely intended for humans rather than machines or automated agents.

### 2.2 Relevant Cyberinfrastructure Support

This section expands on Scientific Workflow Management Systems (SWFMS) and its significance for the design of GEO-SEED. The second section discusses ontologies and their significance in the GEO-SEED design.

**Scientific Workflow Management Systems**

Computational entities range from stand-alone atomic software to composite combinations of software and data, to workflows, used to accomplish a scientific task to data sets, picture files, and maps. The business application domain has a long history of capturing their processes as workflows. The *Workflow Management Coalition* defines workflow as “The automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules and have used the technology to support their endeavors” (WFMC, 1999). Scientists are now moving in that direction to describe their processes. The *Workflow Management Coalition* defines a Workflow Management System (WFMS) as “A system that defines, creates and manages the execution of workflows through the use of software, running on one or more workflow engines, which is able to interpret the process definition, interact with workflow participants and, where required, invoke the use of IT tools and applications” (WFMC, 1999). To further automate that process, researchers are adapting
those concepts into Scientific Workflows and Scientific Workflow Management Systems (Ludäscher, et al., 2006).

A SWFMS, Scientific Workflow Management System, is a system that supports the specification, modification, run, re-run, and monitoring of a scientific workflow using the workflow logic to control the order of executing workflow tasks (Ludäscher, 2006) (Chebotko, et al., 2007) (Lin, et al., 2008). This type of system maintains metadata about workflow tasks, their implementation, and invocation in some type of a repository (Ludäscher, et al., 2006), (Oinn, 2006), (Scheidegger, 2008), (Churches, 2006), (Deelman, 2005), (Zhao, 2007), (Lin, 2009), (Wibisono, 2007). An example of a task manager can be found in (Lin, 2009). Currently, scientists have to register or search for a Web service in a specialized metadata repository and then re-register it with a SWFMS, which doubles their effort. Machine-interpretable metadata is one step towards the automation of this process. Work has been done exploring efficient storage and querying of such metadata using relational databases (Chebotko, 2007) (Chebotko, 2009).

More and more scientists have begun to use workflow technologies to automate the steps of a scientific process, ranging from collecting raw datasets to creating new products that can lead to scientific discovery. Defining the steps to create a final data product can be streamlined by SWFMS’s, which support the specification, execution, re-run, and monitoring of computational entities (Ludäscher, 2007), (Deelman, 2005), (Gates, 2007), (Ludäscher, et al., 2006), (Battle, 2005), (Churches, 2006), (Sahoo, 2008), (De Roure, 2009). Finally, databases that record scientific workflow provenance and monitoring information (Simmhan, 2006), (Simmhan, 2005), (Zhao, 2008), (Davidson, 2007), (Kim,
2008), (Chebotko, 2007), (Valerio, 2008), (Cruz, 2008) are of value to work described in this dissertation.

**Ontologies**

Ontologies play a key role in the sharing and discovery of information found on the Web and remain important components in the development and implementation of the Semantic Web and Web2.0. Gruber gave the following short definition of an ontology that has been widely accepted as follows: “ontology is a specification of a conceptualization”. The article goes on to state: “We use common ontologies to describe *ontological commitments* for a set of agents so that they can communicate about a domain of discourse without necessarily operating on a globally shared theory. Pragmatically, a common ontology defines the vocabulary with which queries and assertions are exchanged among agents. In short, a commitment to a common ontology is a guarantee of consistency with respect to queries and assertions using the vocabulary defined in the ontology” (Ontology, 1993).

Further Gruber discusses ontologies and their relationship to the Semantic Web and their key applications. Ontologies are part of the W3C standards used in the Semantic Web to specify standard conceptual vocabularies. This allows for data schema representation that can be exported, translated, queried, and unified across independently developed systems and services. Relevant applications to date include database interoperability, cross database search, and the integration of web services (Gruber, 2009).
2.3 Semantic Web Technologies

Semantic Web Overview

At its core, the Semantic Web comprises a set of design principles, collaborative working groups, and a variety of enabling technologies. One application of the Semantic Web is to enable users to locate, select, employ, compose, and monitor Web-based services automatically (Shadbolt, 2006).

*Semantic Web* is a term coined by World Wide Web Consortium (W3C) director Sir Tim Berners-Lee. It describes methods and technologies to allow machines to understand the meaning or "semantics" of information on the World Wide Web.

According to the original vision, the availability of machine-readable metadata would enable automated agents and other software to access the Web more intelligently. The agents would be able to perform tasks automatically and locate related information on behalf of the user (Semantic_Web, 2009).

The Semantic Web is a network of data described and linked in ways to establish context or semantics that adhere to defined grammar and language constructs. (Hebeler, 2009). The Semantic Web consists of a set of flexible statements of many types that allow for the formation of rich expressions, allow for simplified integration and sharing, enabling inference, and extraction of meaningful information.
The Semantic Web is made up of relationships, which include definitions, associations, aggregations, and restrictions. Statements and corresponding relationships establish concepts and instances, together which form an ontology.

Major programming components for a Semantic Web application include Statement, URI, Language, Ontology, and Instance Data.

**Semantic Web Languages - RDF, SPARQL, OWL-S and SWSO**

RDF (Resource Description Framework), which is accessible for queries and integration, is a standard accepted by W3C to standardize information stored on the Web. RDF uses a directed, labeled graph data format for representing information on the Web. The graph data format is also represented as a triple, which consists of a subject, predicate, and object (s, p, o), where the predicate gives the relationship between the subject and the object. This triple format supports storing the RDF data in a database (RDF, 2004).

SPARQL is the query language for RDF that can be used to express queries across diverse data sources (SPARQL, 2008). To make use of a Web service, a software agent needs a computer-interpretable description of the service, and the means by which it is accessed. An important goal for Semantic Web markup languages, then, is to establish a framework within which these descriptions are made and shared.

Web sites should be able to employ a standard ontology, consisting of a set of basic classes and properties, for declaring and describing services. The ontology structuring mechanisms of OWL (Ontology Web Language) and OWL-S (Semantic Markup for Web
Services) provide an appropriate, Web-compatible representation language framework within which to do this (OWL-S, 2006).

Web Discovery Ontologies exist in several areas related to this work. The Semantic Web Service Ontology (SWSO) (SWSO, 2005) (Battle, 2005) describes a set of service descriptors suitable for storing information about a web service, including identifying information along with subjective notions of quality of service information. This set of service descriptors was derived predominately from the OWL-S profile (OWL-S, 2006), which builds on OWL. The Semantic Web Service Ontology includes 14 service descriptors that provide basic information about a Web Service. These include Service Name, Service Author, Service Contract Information, Service Contributor, Service Description, Service URL, Service Identifier, Service Version, Service Release Date, Service Language, Service Trust, Service Subject, Service Reliability, and Service Cost. Some of these properties are derived from OWL-S (OWL-S, 2006).

OWL-S Semantic Markup for Web Services supports the need of Web users and software agents to discover, invoke, compose, and monitor Web resources. OWL-S is an ontology of services to enable this. It consists of three main parts: the service profile for advertising and discovering services: the process model, which gives a detailed description of a service operation; and the grounding, which provides details on how to interoperate with a service, via message. The service profile contains the following properties: serviceName, textDescription, and contactInformation. The type of a service listed as either atomic or composite is also included.
Web 2.0

Web 2.0 encompasses an electronic world of social networking and the Semantic Web. According to Wikipedia, “A social network is a social structure made of individuals (or organizations) called nodes, which are tied (connected) by one or more specific types of interdependency, such as friendship, kinship, financial exchange, dislike, sexual relationships, or relationships of beliefs, knowledge or prestige” (Social_network, 2009). In particular, wiki technology can be applied to the challenges of widespread sharing and collaboration in a particular application area enabling the scientist to share their expertise in a collaborative environment that has become familiar and easy to use. “A wiki is a website that uses wiki software, allowing the easy creation and editing of any number of interlinked Web pages, using a simplified markup language or a WYSIWYG text editor, within the browser. Wikis are often used to create collaborative websites, to power community websites, for personal note taking, incorporate intranets, and in knowledge management systems.” (Wik) A wiki has the capability to attach metadata, to support identification of the geoscientist’s data and procedures, through classification (ontologies) for the discovery processes as well as domain specific ontologies to tailor the searches and annotations, using the scientist’s terminology. The wiki stores this information in a format that is accessible for the human user. Some wiki implementations store the information in a format that is available for software systems to process the information, such as a Scientific Workflow Management System (SWFMS). The capability for the machine to process the information is a result of the storage organization for the wiki data. RDF triple storage
(Brickley, 2004) is a possible method to store information to support the capability for a machine to process the information.

**Web Services: WSDL and WSDL-S**

WSDL is an XML format for describing Web Services as a set of endpoints operating on messages containing either document-oriented or procedure-oriented information (Christensen, 2001). WSDL is used in conjunction with SOAP 1.1, HTTP GET/POST, and MIME. WSDL operates at the syntactic level but lacks the semantic expressivity needed to represent the requirements and capabilities of Web services. WSDL-S, Web Services Semantics, attempts to solve this problem and includes semantic information such as precondition, input, output and effects of Web service operations (WSDL-S, 2005).

**2.4 Related Work**

**Scientific Knowledge Management using Technologies of the Semantic Web and Social Sharing**

The most closely related work that uses Semantic Web paradigms for scientific knowledge management and sharing are myExperiment (De Roure, 2009) (myExperiment, 2007) and BioCatalogue (BioCatalogue, 2008), led by David De Roure and Carole Goble. myExperiment focuses on building virtual research communities to support the sharing and use of scientific workflows. myExperiment design extends the methods to access information that is currently stored in existing portals, which is closely related to the GEO-SEED design goals. However, myExperiment has focused mostly on the life science domain, but has been extended to the Music and other domains. BioCatalogue also uses wiki technology and
provides the capability to register, annotate, and search for Web Services in the biology and life sciences. While BioCatalogue, myExperiment, and GEO-SEED promote a similar idea to provide a single registration point for web services and enable the creation and sharing of scientific workflows, the projects target different domains – life and earth sciences.

Similar to GEO-SEED, BioCatalogue utilizes an ontology based on myGrid Ontology (myGrid, 2008) to describe the bioinformatics research domain and the dimensions with which a service can be characterized from the perspective of the scientist and is logically separated into two distinct components, the service ontology and the domain ontology. The domain ontology acts as an annotation vocabulary including descriptions of core bioinformatics data types and their relationships to one another, and the service ontology describes the physical and operational features of web services, such as inputs and outputs.

myGrid tools are being used by researchers in a large number of different domains varying from astronomy to medical imaging to music to database and service support to allow novel research or to speed up, formalize or share existing research. The primary tool, Taverna Workbench (Taverna, 2010) is an open source tool for designing and executing workflows created by the myGrid project and funded through the OMII-UK (myGrid, 2008).

**Semantic Web Service Discovery**

Work has been done in Semantic Web Service Discovery Methods. Semantic Web Services (SWS) are summarized by Ngan, Kirchberg, and Kanagasabai (Ngan, 2010). OWL-S uses service profiles and domain ontologies to decide whether there is a match between a requested service and an advertised service Paolucci, et al, (Paolucci, 2002), Klusch
WSMO or Web Service Modeling Ontology - based approaches view service discovery as a task of fulfilling goals which are abstractions of user’s desires, which supports the deployment and interoperability of Semantic Web Services. It has four main components: Goals, Ontologies, Mediators providing interoperability between different ontologies, and Web services. Related work can be found in (Keller, 2005), (Brambilla, 2009), (Vitvar, 2009) (Klusch, 2009a).

SAWSDL and WSDL-S-based approaches enable semantic annotations for web services by using and building on the existing extensible framework of WSDL. Some examples can be found at (Verma, 2005), (Oundhakar, 2005), (Kourtesis, 2008), (Klusch, 2009b). RESTful services need semantic descriptions to facilitate automated discovery and Web 2.0 are being explored to exploit social tagging to enhance service discovery (Ngan, 2010).

**Database management for RDF**

Most existing RDF stores use a Relational Database Management System (RDBMS) as a backend to manage RDF data (Chebotko, 2007). Chebotko’s, work presents a semantics preserving SPARQL-to-SQL translation. This sets the background work to allow discovery of RDF data that is stored in a traditional relational database. The SPARQL-to-SQL translation is shown to allow efficient SPARQL queries for triple RDF stores in SQL databases (Chebotko, 2009). Lin, Lu, Lai, Chebotko, Fei, Hua, and Fotouhi have developed a Visual Scientific Workflow Management System, VIEW (Lin, et al., 2008) (Chebotko, et al.,
2007), which will be explored as an extension to the work to realize using RDF stores to automate the formation of scientific workflows.
CHAPTER 3
SCIENTIFIC COMPUTATIONAL ENTITY DISCOVERY ONTOLOGY

The motivation for using an ontology-based approach for annotation of computational entities and discovery is twofold. On the one hand, searching for computational entities such as Web services and other kinds of software requires metadata that describes instances of computational entities. One of the major problems encountered in the sharing of such metadata is the complexity of processing queries written in natural language due to non-standardized terminology and inconsistent terms for similar or identical entities. An ontology addresses this problem by providing a standard vocabulary for such metadata acquisition. On the other hand, an ontology can be represented in Semantic Web languages, such as OWL and RDFS, which can further be interpreted by machines or automated agents to enable automatic Web services discovery. This chapter describes the work done to create the Computational Entity Discovery Ontology through the development of use cases, the concepts of the ontology, including the description of scientific computational entities, profiles, and ontologies, and the relationships among the concepts.

3.1 Use Cases

This section contains a Use Case Diagram (Figure 1) of the use cases associated with geoscientists who employ computational entities in their work. The diagram in is followed by more detailed descriptions of the actors and actions that make up the use cases.
Figure 1: Geoscience Use Case Diagram
## Use Case Descriptions

<table>
<thead>
<tr>
<th>Name</th>
<th><strong>Share Scientific Entity</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Actors</td>
<td>Geoscientist, Software Developer</td>
</tr>
</tbody>
</table>
| Steps | 1. Geoscientist is engaged in new research area.  
2. Needs to research related areas for the research, so determines related key terms.  
3. Searches and discovers GEO-SEED.  
4. After registering in GEO-SEED, searches for desired key terms.  
5. Inspects related works found.  
6. Determine which profiles are necessary to share the work. If new software fill in the Invocation profile. Fill in the General information profile. If it is a new computational entity, fill in the appropriate information in the Implementation profile. If you are entering a software computational entity, fill in the Invocation profile. If your software is available for download from the Web, enter the deployment information in the Deployment profile. Fill in any Quality of Service information that can be shared. Fill in any related ontology information.  
7. Save and view the entry.  
8. Logout. |

<table>
<thead>
<tr>
<th>Name</th>
<th>Discover Scientific Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actors</td>
<td>Geoscientist or Student</td>
</tr>
<tr>
<td>Steps</td>
<td>1. Login to GEO-SEED.</td>
</tr>
<tr>
<td></td>
<td>2. Enter search information.</td>
</tr>
<tr>
<td></td>
<td>3. View information about the computational entity.</td>
</tr>
<tr>
<td></td>
<td>4. If the information is what you need, determine if you need to</td>
</tr>
<tr>
<td></td>
<td>deploy the entity to your computer and follow directions to</td>
</tr>
<tr>
<td></td>
<td>deploy. If the entity is an application, follow the directions</td>
</tr>
<tr>
<td></td>
<td>to invoke the application, noting required inputs and noting</td>
</tr>
<tr>
<td></td>
<td>defined effects.</td>
</tr>
<tr>
<td></td>
<td>5. If the search results in more than one result, inspect the</td>
</tr>
<tr>
<td></td>
<td>general information profile and the Quality of Service profile</td>
</tr>
<tr>
<td></td>
<td>for information to assist you in deciding which one meets your</td>
</tr>
<tr>
<td></td>
<td>needs.</td>
</tr>
<tr>
<td></td>
<td>6. Follow links to use the desired entity.</td>
</tr>
<tr>
<td></td>
<td>7. After using, you may enter Quality of Service descriptors.</td>
</tr>
<tr>
<td></td>
<td>Save and Exit.</td>
</tr>
<tr>
<td></td>
<td>8. Logout.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Share Data Set Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actors</td>
<td>Geoscientist</td>
</tr>
<tr>
<td>Steps</td>
<td>1. Login to GEO-SEED.</td>
</tr>
<tr>
<td></td>
<td>2. Fill out information about the data set in the provided forms.</td>
</tr>
</tbody>
</table>

24
<table>
<thead>
<tr>
<th>Name</th>
<th>Rate Scientific Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actors</td>
<td>Geoscientist or Student</td>
</tr>
</tbody>
</table>
| Steps | 1. Login to GEO-SEED.  
2. After using a computational entity found in GEO-SEED, you rate the scientific entity to share your experience. Retrieve the Quality of Service profile, and fill in appropriate descriptor fields, checking the required data types.  
3. Add a general rating and a user review for the computational entity for future users to learn of your experience.  
4. Follow links to use the desired entity |

<table>
<thead>
<tr>
<th>Name</th>
<th>Deploy Scientific Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actors</td>
<td>Geoscientist or Student</td>
</tr>
</tbody>
</table>
| Steps | 1. Login to GEO-SEED.  
2. Enter search information.  
3. View deployment information (how to download software to your computer).  
4. Follow links to deploy the desired entity. |
<table>
<thead>
<tr>
<th>Name</th>
<th><strong>Invoke Scientific Entity</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Actors</td>
<td>Geoscientist or Student</td>
</tr>
<tr>
<td>Steps</td>
<td>1. Login to GEO-SEED.</td>
</tr>
<tr>
<td></td>
<td>2. Enter information to search for desired software.</td>
</tr>
<tr>
<td></td>
<td>3. View information about the computational entity software.</td>
</tr>
<tr>
<td></td>
<td>4. Follow links to invoke (run) the software.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th><strong>Automatic Creation of new workflow [Future Implementation]</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Actors</td>
<td>Geoscientist, SWFMS</td>
</tr>
<tr>
<td>Steps</td>
<td>1. Geoscientist requests new workflow.</td>
</tr>
<tr>
<td></td>
<td>2. Geoscientist identifies entities to be included in the workflow.</td>
</tr>
<tr>
<td></td>
<td>3. SWFMS creates the workflow and enters it into GEO-SEED.</td>
</tr>
</tbody>
</table>

### 3.2 The Scientific Computational Entity Discovery Ontology Concept

While the design of our Scientific Computational Entity Discovery Ontology facilitates the reuse of existing efforts on services annotation, such as Semantic Markup for Web Services (OWL-S, 2006), Semantic Web Services Ontology (SWSO) (OWL-S, 2004) (Battle, 2005), Web Services Description Language (WSDL) (Chinnici, 2007), and Web Service Semantics (WSDL-S) (WSDL-S, 2005), it also features additional Web services descriptors that avoid the complexity of OWL-S and can be used by moderately prepared users. At the heart of the discovery ontology is the notion of a *Computational Entity* that models any computational
entity used in scientific research. For the purposes of this discussion, the term Web Service is used to identify any computational entity that can be found on the Web. When identifying types of computational entities, the term web service is used as a possible type and refers to the accepted technical definition of an entity implemented as a web service. A computational entity can have multiple implementation types, such as application, composite application, web service, workflow, process, data store, ontology, other.
Figure 2: Scientific Computational Entities Discovery Ontology Graph

Figure 2: Scientific Computational Entities Discovery Ontology Graph is a graph, which illustrates most of the concepts and their relationships in our Scientific Computational Entities Discovery Ontology. Many of the shown leaf nodes represent ranges of properties that are not specified within the ontology, but can be restricted to classes in some other ontology.
For example, an author (author descriptor) may be a string literal or an instance of a person described using the FOAF (Friend of a Friend) ontology (Miller, 2007).

### 3.3 The Ontology Profile Concept

In the discovery ontology presented in this work, the descriptors of a Web service, computational entity that is software related, are organized into six profile categories: `GeneralProfile`, `QoSProfile`, `InvocationProfile`, `DeploymentProfile`, `ImplementationProfile`, and `GeoscienceProfile`. The six categories were chosen to support the kind of information that one would need to know to find, use, and rate a scientific computational entity. Not all categories and descriptors within those categories will have information to enter. This section includes an object diagram (Figure 3: GEO-SEED Scientific Computational Entity Discovery Ontology) showing the Discovery Ontology Concept with the relationships of Scientific Computational Entity to the profile categories. It also includes tables showing the Profile descriptors, purpose and type for each descriptor.
Figure 3: GEO-SEED Scientific Computational Entity Discovery Ontology
The following table shows the profile names and purpose, followed by tables listing the descriptors for each profile with a description and data type for each descriptor.

**Table 1: Profiles of Scientific Computational Entities**

<table>
<thead>
<tr>
<th>PROFILE NAME</th>
<th>PURPOSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeneralProfile</td>
<td>Basic Identifying Information</td>
</tr>
<tr>
<td>QoSProfile</td>
<td>Quality-of-service characteristics</td>
</tr>
<tr>
<td>InvocationProfile</td>
<td>Details needed to execute an entity</td>
</tr>
<tr>
<td>DeploymentProfile</td>
<td>Details needed to download and deploy an entry from the Web</td>
</tr>
<tr>
<td>ImplementationProfile</td>
<td>Information related to the implementation of a computational entity</td>
</tr>
<tr>
<td>GeoscienceProfile</td>
<td>Information obtained with a domain specific ontology</td>
</tr>
</tbody>
</table>
Table 2 - Table 7 present the descriptors for each profile category and provide the data type for each descriptor.

The *GeneralProfile* descriptors (Table 2) capture basic information about a Web service, including its name, authors, contact information, contributors, description, location, unique identifier, version, release date, ISO language, subject, type, cost, license, and support. Many of these descriptors are reused from the SWSO ontology (SWSO, 2005). *GeneralProfile* is related to Scientific Computational Entity via the subject property.
Table 2: GeneralProfile Descriptors

<table>
<thead>
<tr>
<th>DESCRIPTOR</th>
<th>DESCRIPTION</th>
<th>DATATYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Name of entity</td>
<td>string</td>
</tr>
<tr>
<td>Authors</td>
<td>Authors of entity</td>
<td>string list</td>
</tr>
<tr>
<td>Contact Information</td>
<td>Current name or email to contact if needed</td>
<td>string</td>
</tr>
<tr>
<td>Contributors</td>
<td>Names of other contributors</td>
<td>string list</td>
</tr>
<tr>
<td>Description</td>
<td>Brief description of purpose of entity</td>
<td>string</td>
</tr>
<tr>
<td>URIs</td>
<td>URL where entity can be located for use</td>
<td>url</td>
</tr>
<tr>
<td>Unique Identifier</td>
<td>Identifier of entity for search purposes</td>
<td>uri</td>
</tr>
<tr>
<td></td>
<td>example: tawsl08</td>
<td>example: 3.1.4</td>
</tr>
<tr>
<td>Version</td>
<td>Version being registered</td>
<td>date</td>
</tr>
<tr>
<td>Release Date</td>
<td>Date of version release</td>
<td>date</td>
</tr>
<tr>
<td>Languages</td>
<td>Natural Language used</td>
<td>string list</td>
</tr>
<tr>
<td>Subject</td>
<td>Subject in which the entity will be used, e.g. gravity, map, volcano, general.</td>
<td>string</td>
</tr>
<tr>
<td>Type</td>
<td>Examples: application, composition of applications, web service, portal, pictures, dataset</td>
<td>string</td>
</tr>
<tr>
<td>Cost</td>
<td>Cost for using the entity</td>
<td>$xx.xx</td>
</tr>
<tr>
<td>License</td>
<td>Description of any license required</td>
<td>string</td>
</tr>
<tr>
<td>Support</td>
<td>Contact information for support</td>
<td>url</td>
</tr>
</tbody>
</table>
QoSProfile (Table 3) is intended for scientists to collaboratively rate different quality-of-service characteristics of a Web service. This feature is similar to the one that customers, who purchase merchandise online, use to rate and review purchased products by means of comments and ratings on a given scale. Such metadata becomes extremely useful when a scientist has multiple alternatives for implementing the same computational entity. Quality-of-service descriptors include trust, reliability, and availability, processing time, requests per seconds, security, known failures, overall rating, and user reviews.

Table 3: OoS Profile Descriptors

<table>
<thead>
<tr>
<th>DESCRIPTOR</th>
<th>DESCRIPTION</th>
<th>DATATYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trust</td>
<td>Rate dependability- 5 High 0 Low</td>
<td>0..5</td>
</tr>
<tr>
<td>Reliability</td>
<td>Probability system will deliver as expected</td>
<td>0.0..1.0</td>
</tr>
<tr>
<td>Availability</td>
<td>Probability system will be up and running at a given time</td>
<td>0.0..1.0</td>
</tr>
<tr>
<td>Processing Time</td>
<td>Average processing time</td>
<td>time and units</td>
</tr>
<tr>
<td>Requests per Second</td>
<td>Average requests per second software can handle</td>
<td>integer</td>
</tr>
<tr>
<td>Custom Metric</td>
<td>Description of the metric, rating scale, and rating</td>
<td>string , string, and 1..5</td>
</tr>
<tr>
<td>Security</td>
<td>Description of how likely system resists intrusions, 5 High security 0 Unknown</td>
<td>0..5</td>
</tr>
<tr>
<td>Known Failures</td>
<td>Description and reason for known failures</td>
<td>string</td>
</tr>
<tr>
<td>Overall Rating</td>
<td>Rating for the entity, 5 highest</td>
<td>1..5</td>
</tr>
<tr>
<td>User Reviews</td>
<td>Comment on entity from user</td>
<td>string</td>
</tr>
</tbody>
</table>
InvocationProfile (Table 4) contains sufficient details to execute a Web service, which is important not only for geoscientists, but also for tools like scientific workflow management systems. This includes information about the processes and operations, required preconditions, parameters, messages, bindings, and expected effects. While this information is intended for humans, specifications for the linked WSDL and OWL-S urls are used by automated agents.

Table 4: Invocation Profile Descriptors

<table>
<thead>
<tr>
<th>DESCRIPTOR</th>
<th>DESCRIPTION</th>
<th>DATATYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processes and Operations</td>
<td>List of included processes</td>
<td>string list</td>
</tr>
<tr>
<td>Types</td>
<td>Data types of inputs</td>
<td>string list</td>
</tr>
<tr>
<td>Parameters and Messages</td>
<td>List of required parameters</td>
<td>string list</td>
</tr>
<tr>
<td>Preconditions</td>
<td>Conditions that must be true before invocation</td>
<td>string list</td>
</tr>
<tr>
<td>Effects</td>
<td>Postconditions that must be true after completion</td>
<td>string list</td>
</tr>
<tr>
<td>Bindings</td>
<td>soap, xml or other bindings</td>
<td>url</td>
</tr>
<tr>
<td>WSDL Document</td>
<td>Pointer to the WSDL document for Web Service, if applicable</td>
<td>url</td>
</tr>
<tr>
<td>OWL-S Document</td>
<td>Pointer to the OWL-S, if applicable</td>
<td>url</td>
</tr>
</tbody>
</table>
DeploymentProfile (Table 5) is helpful when the user decides to download and deploy a Web service on a local site. The descriptors include information on downloads, software components, installation process, supported operating systems, software dependencies, required hardware, activation and deactivation procedures.

Table 5: Deployment Profile Descriptors

<table>
<thead>
<tr>
<th>DESCRIPTOR</th>
<th>DESCRIPTION</th>
<th>DATATYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>URLs</td>
<td>Location for download source</td>
<td>string</td>
</tr>
<tr>
<td>Software Architecture</td>
<td>Architecture type of implementation, such as 3-tiered, pipeline, client-server</td>
<td>string list</td>
</tr>
<tr>
<td>Operating Systems</td>
<td>Operating systems that are supported</td>
<td>string list</td>
</tr>
<tr>
<td>Software Dependencies</td>
<td>Any software that must be invoked</td>
<td>string list</td>
</tr>
<tr>
<td>Hardware</td>
<td>Hardware platforms supported</td>
<td>string</td>
</tr>
<tr>
<td>Installation</td>
<td>Directions to install</td>
<td>string</td>
</tr>
<tr>
<td>Activation and Deactivation</td>
<td>Direction to activate and deactivate</td>
<td>string</td>
</tr>
</tbody>
</table>

ImplementationProfile (Table 6) capture metadata related to implementation of a Web service, such as source repository location, development environment, programming language, algorithms, libraries, and documentation. This profile supports software developers who work with geoscientists and are required to modify existing Web services to satisfy scientist’s needs.
Table 6: Implementation Profile Descriptors

<table>
<thead>
<tr>
<th>DESCRIPTOR</th>
<th>DESCRIPTION</th>
<th>DATATYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Location of source code</td>
<td>string or url</td>
</tr>
<tr>
<td></td>
<td>Example: Visual Studio, Eclipse</td>
<td>string</td>
</tr>
<tr>
<td></td>
<td>Programming language used</td>
<td>string</td>
</tr>
<tr>
<td></td>
<td>Any known algorithm used in the design</td>
<td>string list</td>
</tr>
<tr>
<td></td>
<td>Classes, libraries, etc</td>
<td>string list</td>
</tr>
<tr>
<td></td>
<td>User manual, source documentation conventions</td>
<td>string list</td>
</tr>
</tbody>
</table>

GeoscienceProfile (Table 7) links scientific processes and their implementations to concepts from a geoscience domain-specific ontology, such as the Gravity Data Ontology (Gates, 2007), via the annotation property (similar to how WSDL-S associates domain semantics to WSDL elements). This way, scientists can use geosciences terminology for registering and discovering Web services.

Table 7: Geoscience Profile Description

<table>
<thead>
<tr>
<th>DESCRIPTOR</th>
<th>DESCRIPTION</th>
<th>DATATYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontologies</td>
<td>Domain-specific ontologies related to entity</td>
<td>string</td>
</tr>
<tr>
<td>Ontology Dependent Annotations</td>
<td>Domain-specific semantic annotations</td>
<td>string list</td>
</tr>
</tbody>
</table>
3.4 Relationship Structure in the Ontology

At the heart of the structure of the discovery ontology is the concept of a Scientific Computational Entity. A computational entity may consist of an application that can be invoked from the Web, a composition of applications, a Web service, a workflow, or a description of a process used by a scientist.

Figure 4 shows several instances of Scientific Computational Entities stored in GEO-SEED. A geoscientist may wish to create a Contour Map that requires a Gridding Process. Several Gridding Entities may have been stored in GEO-SEED, which are available for discovery. Information about the Contouring Process may also be stored as an Entity in GEO-SEED.

![Diagram of Computational Entities](image)

**Figure 4: Instances of Computational Entities**
CHAPTER 4
GEOSCIENCES WEB SERVICES METADATA MANAGEMENT AND DISCOVERY SYSTEM

This chapter describes the GEO-SEED architecture and the motivation for using the wiki technology to support its implementation as a prototype for the geoscientist. It further describes the GEO-SEED Repository organization, possible designs for the organization, and a description of possible queries a geoscientist may make.

A short on-line five question survey was carried out to determine how researchers currently share and discover scientific computational entities they need for their research. Invitations were sent via email to faculty from a randomly selected set of 10 institutions with Ph. D. programs in geology. Faculty members were selected from a cross-section of ranks. Responses were returned by 50% of the invitations. A copy of the survey and responses is included in Appendix B. Research areas were reported in Earth Sciences, Quaternary Geology, Exploration Seismology, Planetary Science, and Geosciences. Researchers were asked to identify what methods they currently use to discover the computational entities to support their research. The researchers reported primarily using professional contacts to discover computational entities (Figure 5). Second most cited source was online search engines. Google was cited by all with Yahoo also cited by two researchers as the search engine they used. They were asked how they currently share their research results. Most reported sharing their results via standard professional presentations and publications and personal Web pages (Figure 6). In the “Other” category for this question, one researcher reported using the first three methods and another
researcher reported using software repositories but did not identify any. This explains the 40% other category showing in the figure.

![Diagram of Discovery of Computational Entities]

**Figure 5: Researcher's Discovery Methods**

In response to the question, “Are you aware of any existing comprehensive Web support system that enables geoscientists to share and discover computational entities?”, while two researchers identified web sites to share software and datasets related to their research area, no one reported a comprehensive on-line site similar to the proposed GEO-SEED wiki.
One researcher reported support at websites http://www.orfeus-eu.org/ and http://www.iris.edu/software/. Another reported support found at geodynamics.org. Descriptions of the support found at these sites follow:

Computational Infrastructure for Geodynamics (CIG) is a membership-governed organization that supports and promotes Earth science by developing and maintaining software for computational geophysics and related fields. They are currently working with software in several sub-disciplines, including mantle convection, short and long time-scale tectonics, computational seismology, and the geodynamo. Plans call for expanding into magma migration.

ORFEUS (Observatories and Research Facilities for European Seismology), is the non-profit foundation that aims at coordinating and promoting digital, broadband (BB) seismology in the European-Mediterranean area. Seismic Data Portal provides access to event, waveform and acceleration data from the EMSC, ORFEUS (VEBSN and EIDA) and a number of acceleration networks.

IRIS distributes seismology software for data requests, visualization and analysis. IRIS is a consortium of over 100 US universities dedicated to the operation of science facilities for the acquisition, management, and distribution of seismological data.

None of these are as comprehensive as the capabilities proposed for GEO-SEED.

4.1 GEO-SEED Architecture

The GEO-SEED architecture is presented in Figure 7. The architecture involves two main components called Web 2.0 Structured Wiki and Semantic Web and Machine-Processable
Knowledge (Knowledge Management System). The wiki provides an ontology-driven collaborative environment for geoscientists to share their knowledge about available Web services and software entities used in geosciences research. The wiki user interface is generated and structured based on the discovery ontology, such that scientists are provided with the means to assign values to each descriptor/property from the ontology. All annotations entered by scientists are stored into the wiki database, which also maintains complete provenance of the annotation process and may be represented by a relational database management system (RDBMS). While the wiki component is solely intended for collecting and managing Web computational entity annotations in the human-friendly form via a recognizable and user-friendly graphical Web-based interface, its ontology-driven information structuring prepares grounds for the second component – Knowledge Management System.
The overall goal of the Knowledge Management System is in acquiring and managing Web service annotations in the machine-interpretable form, enabling scientific workflow management systems and other automated agents to discover computational resources that can be used to implement workflow tasks and sub-workflows, aiding in semi-
automatic workflow design and composition. The system is responsible for extracting information from the wiki and representing it in a format that can be processed by a machine, such as OWL and RDF. Discovery and domain ontologies are stored into an OWL repository and accessed via the ontology interface of the system. Extracted RDF annotations are stored into an RDF repository and accessed via the query/discovery interface that uses SPARQL queries to search for Web service annotations. Both OWL and RDF repositories may be represented by one or two RDF stores.

The GEO-SEED architecture enables the following two sample scenarios for a scientist and scientific workflow management system (SWFMS) interaction with GEO-SEED. In the first scenario, the geoscientist can search GEO-SEED’s wiki to retrieve and update information about existing Web services that support particular scientific computations. Different geoscientists can exchange their experience with a Web service by annotating its various ontological descriptors and commenting on relevant issues. In the second scenario, the SWFMS that is used by the geoscientist for a workflow design may access GEO-SEED to retrieve information about Web services that can be used to implement a workflow computational entity, presenting the scientist with available options. Upon the scientist response, which may involve visiting Web service wiki pages, the SWFMS may use GEO-SEED’s service invocation information to execute the workflow task. Besides these two scenarios, one can envision other numerous use cases of GEO-SEED, such as automatic switching between alternative task implementations for fault tolerance or result verification, automating workflow composition, getting support for a Web service, deploying software locally, and so forth.
4.2 Structured Wiki Implementation

The GEO-SEED wiki was implemented using the XWiki software (XWi08), which served as a generic and extensible platform for developing collaborative applications using the wiki paradigm. XWiki was selected as a second generation structured wiki, allowing information to be collected and presented in a well-organized structured fashion, supporting our long-term goal for integrating these processes into Scientific Workflow Management Systems. XWiki was deployed as Java servlets over the Java Tomcat container and used MySQL as a wiki database. Service annotations were extracted from the database, represented in the RDF format and stored into a relational RDF store for future querying.

APPENDIX A, presents screen shots of the GEO-SEED wiki with data for The Generic Mapping Tool that was found online. The screen shots show the description of the computational entity and the descriptors for GeneralInformation, DeploymentInformation, and ImplementationInformation filled in.

4.3 GEO-SEED RDF Repository

RDF Triple Store Design

While GEO-SEED can employ existing relational RDF stores, such as Jena (Wilkinson, 2003), Sesame (Broekstra, 2002) or RDF-Prov (Chinnici, 2007) (Chebotko, 2007) for storing and querying extracted RDF triples, we notice an interesting property of RDF data acquired with our Scientific Computational Entity Discovery Ontology that motivated our novel design. In particular, most RDF triples in GEO-SEED belong to one of the six profiles and
frequently all the triples for one profile are retrieved by a query; therefore, storing all the triples of one profile in one database table can speedup query evaluation.

Database storage schema’s can be classified into five approaches (Chebotko, 2009) (Brazier, 2010):

- Schema-Oblivious: A single general-purpose relation is used to store all RDF triples.
- Schema-Aware: An ontology, such as WSDO, is used to generate database schemas with relations that correspond to classes and properties in the ontology.
- Data-Driven: Patterns in RDF data are used to generate relations on the fly.
- User-Customizable: Relations are defined by a user via generic mechanism (triple patterns) that describes what triples can be stored in these relations.
- Hybrid: A database schema is generated by any combination of the previous approaches.

<table>
<thead>
<tr>
<th>Table 8: Extracted RDF Triples</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>=&lt;:WS1&gt;&lt;:describedBy&gt;&lt;:GP1&gt; .</code></td>
</tr>
<tr>
<td><code>=&lt;:WS1&gt;&lt;:describedBy&gt;&lt;:QoSP1&gt; .</code></td>
</tr>
<tr>
<td><code>=&lt;:GP1&gt;&lt;rdf:type&gt;&lt;:GeneralProfile&gt; .</code></td>
</tr>
<tr>
<td><code>=&lt;:QoSP1&gt;&lt;rdf:type&gt;&lt;:QoSProfile&gt; .</code></td>
</tr>
<tr>
<td><code>=&lt;:GP1&gt;&lt;:subject&gt;&lt;:Gridding&gt; .</code></td>
</tr>
<tr>
<td><code>=&lt;:GP1&gt;&lt;:author&gt;&quot;Pearl Brazier&quot; .</code></td>
</tr>
<tr>
<td><code>=&lt;:QoSP1&gt;&lt;:trust&gt;&quot;3&quot; .</code></td>
</tr>
<tr>
<td><code>=&lt;:QoSP1&gt;&lt;:availability&gt;&quot;0.9&quot; .</code></td>
</tr>
<tr>
<td><code>=&lt;:QoSP1&gt;&lt;:overallRating&gt;&quot;4&quot; .</code></td>
</tr>
</tbody>
</table>

Many existing relational RDF stores use the single table schema-oblivious approach to store all triples in an RDF dataset. For example, Table 8 shows some extracted RDF triples.
for a Web service that has been registered in the GEO-SEED wiki. Table 9 shows how these triples can be stored in a single database table, Triple(s, p, o).

**Table 9: Triple Table**

<table>
<thead>
<tr>
<th>triple</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
</tr>
<tr>
<td>p</td>
</tr>
<tr>
<td>o</td>
</tr>
<tr>
<td>:WS1 rdf:type</td>
</tr>
<tr>
<td>:WS1 describedBy</td>
</tr>
<tr>
<td>:WS1 describedBy</td>
</tr>
<tr>
<td>:GP1 rdf:type</td>
</tr>
<tr>
<td>:QoSP1 rdf:type</td>
</tr>
<tr>
<td>:GP1 subject</td>
</tr>
<tr>
<td>:GP1 author</td>
</tr>
<tr>
<td>:QoSP1 trust</td>
</tr>
<tr>
<td>:QoSP1 availability</td>
</tr>
<tr>
<td>:QoSP1 overallRating</td>
</tr>
</tbody>
</table>

The Triple table and its indexes can become very large over time affecting the efficiency of query evaluation, which may require self-joins of the table. A popular approach to resolve this problem is to partition the table horizontally into several schema-aware tables based on the value of a triple predicate (column p) (Wilkinson, 2006) (Godbole, 2008). Resulting tables are called Property tables; three of them are shown as follows: type, trust, and availability.
Table 10: Property Table for type

<table>
<thead>
<tr>
<th>type</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>o</td>
</tr>
<tr>
<td>:WS1         :WebService</td>
<td></td>
</tr>
<tr>
<td>:GP1         :GeneralProfile</td>
<td></td>
</tr>
<tr>
<td>:QoSP1       :QoSProfile</td>
<td></td>
</tr>
</tbody>
</table>

Table 11: Property Table for trust

<table>
<thead>
<tr>
<th>trust</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>o</td>
</tr>
<tr>
<td>:QoSP1        :3</td>
<td></td>
</tr>
</tbody>
</table>

Table 12 Property Table for availability

<table>
<thead>
<tr>
<th>availability</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>o</td>
</tr>
<tr>
<td>:QoSP1        :0.9</td>
<td></td>
</tr>
</tbody>
</table>

Predicate-based partitioning yields smaller tables and indexes. Moreover, it saves space since predicate values are encoded as table names rather than column values; however, with this design, retrieving all triples in one profile requires computing the union of multiple property tables, which is not desirable as it may increase and randomize disk access operations.
Our design is also based on the idea of horizontal partitioning. But this time, table Triples is partitioned into multiple user-customizable tables, where each table stores all triples that belong to one of the six profiles defined in our ontology. These tables are called Profile tables. For example, tables for GeneralProfile and QoSProfile are as follows:

Table 13: Profile Table for GeneralProfile

<table>
<thead>
<tr>
<th>GeneralProfile</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
</tr>
<tr>
<td>:GP1</td>
</tr>
<tr>
<td>:GP1</td>
</tr>
<tr>
<td>:GP1</td>
</tr>
</tbody>
</table>

Table 14: Profile Table for QoSProfile

<table>
<thead>
<tr>
<th>QoSProfile</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
</tr>
<tr>
<td>:QoSP1</td>
</tr>
<tr>
<td>:QoSP1</td>
</tr>
<tr>
<td>:QoSP1</td>
</tr>
<tr>
<td>:QoSP1</td>
</tr>
</tbody>
</table>

Triples that do not belong to any profile can still be stored using property tables.
To better compare the three presented relational RDF store designs, consider the SPARQL query that retrieves the quality-of-service descriptors of a Web service :WS1:

```
Select ?profile ?pre ?obj
Where {
  :WS1 :describedBy ?profile .
  ?profile rdf:type :QoSProfile .
}
```

To evaluate this query for the case of the single table approach, two joins need to be computed:

```
Triple ⊗ Triple ⊗ Triple.
```

For the property tables approach, either two joins:

```
describedBy ⊗ type ⊗ (trust ∪ reliability ∪ availability ∪ ⋯ ∪ userReview)
```

or many joins can be evaluated:

```
(describedBy ⊗ type ⊗ trust) ∪ (describedBy ⊗ type ⊗ reliability) ∪ (describedBy ⊗ type ⊗ availability) ∪ ⋯ ∪ (describedBy ⊗ type ⊗ userReview).
```

While the two-join evaluation seems to be more appealing, one should note that the intermediate result, obtained after the union of all the property tables that correspond to the quality-of-service profile descriptors, is not indexed; thus, the join with the intermediate result may be slow. On the other hand, computing many joins allows the use of indexes, but it still does not look promising since the joins between tables `describedBy` and `type` are recomputed many times, wasting computation.
Finally, for the profile tables approach, the query only requires one join:

\[ \textit{describedBy} \bowtie \textit{QoSProfile}. \]

It should be evident that our design results in much better query complexity. The \textit{Triple} table, property tables, and profile tables are readily supported by the generic and semantics-preserving SPARQL-to-SQL translation (Chebotko, 2009) that we employ for the GEO-SEED repository implementation.

An empirical comparison of the three approaches is presented in the next chapter.
CHAPTER 5
EVALUATION AND ASSESSMENT

The evaluation and assessment consisted of three parts. Initially a paper exercise was administered to test the feasibility and concept of overall design of the discovery ontology. Secondly a computerized test of the usability of the GEO-SEED prototype was performed. And thirdly a performance study of the search and discovery capability was carried out. This chapter gives further details of the evaluation and assessment of the three exercises.

5.1 GEO-SEED Usability Study-I

GEO-SEED Usability Study I answers the question: Can a human interact with the system to document information about scientific computational entities?

A usability study of the prototype was conducted with 18 students. The goal of the case study was to demonstrate the feasibility and usability of the wiki dedicated to registration, discovery, and annotation of scientific software.

Description of the study

The study grew out of a scenario in which a geoscientist may want to create a Gravity Contour Map using existing gravity data.

The Gravity Contour Map workflow shown in outlines the required data and the methods needed to create a Gravity Contour Map. This workflow was used as a sample computational entity for the case study. Raw gravity data for the map is gathered from
readings at a base station and readings that come from a gravity station. These gravity readings are merged and converted by Bouguer Anomaly methods. This processed data in turn is input into the gridding process. The output from the gridding process is then input into the Contour process, which produces the desired Contour map. We selected the Gridding method for use in the case study (Gates, 2007) (Salayandia, 2006; Salayandia, 2006).

Figure 8: Gravity Contour Map Workflow
In the study we selected a Web page that presented a software package supporting the Gridding Process. The page was linked to another page describing how to download the software. Participants were asked to provide personal background information and then to identify and fill in the descriptors for the GeneralProfile, QoSProfile and the DeploymentProfile based on the given information. See APPENDIX C for copies of the case study documents.

**Results**

The participants reported having little knowledge of geology, were knowledgeable Internet users, and had knowledge of web applications, had less knowledge about web services, and a minimal knowledge of workflows.

The following were the results of the usability study. Over half of the participants were able to identify more than 70% of the descriptors for the General Info Profile. For the QoS profile, participants were able to identify more than 50% of the descriptors. However, it should be noted, most of this information was not available on the Web pages supplied. For the Deployment profile, participants successfully identified descriptors for an average of 42% of the descriptors. Only two descriptors were found by more than 50% of the participants, but this reflects the limited information that was provided on the Web page. As might be expected, filling in the General Profile was the least difficult of the three profiles presented and also reflects the amount of information that was given.

The participants were given the opportunity to indicate whether they needed help on understanding the meaning of each descriptor. On average only 13% of the participants
indicated that they would need help to understand the meaning of the General Profile descriptors, with UniqueIdentifier requiring the most help. For the QoS profile, on the average 38% of the fields required help, with Custom Metric requiring the most help. Overall rating, user reviews, and reliability descriptors needed the least help. The Deployment profile required help on an average of 29% of the descriptors with Software Dependencies and Activation and Deactivation descriptors requiring the most help.

As a result of the Usability Study I, we came to a conclusion that someone who is a typical user of the Internet would be able to, with little help, use GEO-SEED effectively. Indeed, participants were, in a reasonable amount of time, without extensive help, successfully able to find and enter the information into the General Profile, and the Deployment Profile. Adding QoS information may require more extensive help on the meaning of the descriptors. By monitoring the use of help in the registration process, in the spirit of social computing, the scientific wiki community will also be able to contribute to the future evolution of the discovery ontology.

5.2 GEO-SEED Usability Study-II

GEO-SEED Usability Study II answers the question: Can a human interact with the system to document and retrieve information about scientific computational entities?

Description of study

Following the implementation of GEO-SEED, a more extensive Usability Study was conducted in which participants were given directions to access GEO-SEED, register, and login to the system. They were then directed to register some software, process, portal, or
known dataset. They were then asked to logout, log back in and search for and update information that was previously stored. After completing that part of the exercise, the participants were instructed to add annotations to what they discovered. A copy of the study directions is included in APPENDIX D.

After participating in the Usability exercise, the participants were asked to participate in a ten question online survey of their experience. Copies of the Survey, results and descriptive analysis data are also included in APPENDIX D.

**Descriptive Statistics Results**

Invitations to participate in the Usability Study were sent to 31 Researchers, Faculty, Graduate and Undergraduate Students. Surveys were returned by 17 individuals, for a 55% return rate. A descriptive analysis of the results follows:

**Background information of participants**

33% of participants were Researchers or Professors, 20% were graduate students, and 40% were undergraduate students.

The educational level of the participants was: Ph.D. 66.7%; Masters 18.8%; and undergraduate degrees 37.5%. Some participants selected more than one category, hence sum greater than 100%.

42.9% indicated more than minimal knowledge of the Geology field. Fields of interest identified were quaternary geology, geophysics, structural geology, and mineral chemistry.
All participants were experienced internet users. 66.7% reported themselves as Experienced Technical Users.

Participants were asked about their knowledge of Web related Technologies. Most were quite knowledgeable of Web Service, WSDL, Workflow and Web 2.0, with Knowledge of Workflows being the least knowledgeable at 71% (Figure 9).

![Knowledge of Web Technologies](chart)

**Figure 9: Current Knowledge of Web Related Technologies**

**Assessment of the overall rating for GEO-SEED**

On the final question in which the participants were asked to give an overall rating for GEO-SEED; 40% strongly agree and 33.3% agree that GEO-SEED would be a useful tool for the geoscientist, while 26.7% indicated disagreement (Figure 10). We concluded that GEO-
SEED would be a useful system to support researchers and students, given that 73.3% gave a positive response.

Figure 10: Overall Rating of GEO-SEED Usability

Conclusion #1:

GEO-SEED can support the geoscientist to share their information, however, the user interface to accomplish the desired tasks could be improved.
Detailed analysis of Ontology Features

For the most part, between 40% and 50% of participants indicated no difficulty registering information into the descriptor fields for a computational entity. Between 20% and 35% showed some difficulty in varying degrees (Figure 12).

Conclusion #2:

GEO-SEED’s organization would be a supportive environment with a sufficient range of information to support the geoscientist to share and discover their computational entities.
On performing the task of searching for information stored in GEO-SEED, 80% of participants agreed they had no difficulty. On understanding the meaning of the descriptor fields, over 60% had no difficulty, while over 30% expressed difficulty with this aspect. On finding help, only less than 40% indicated they were able to find help when they needed it (Figure 13).

**Conclusion #3:**

The help features could be improved to support understanding of the GEO-SEED profiles.
When polled about how GEO-SEED may be helpful for the proposed tasks and audiences, between 47% and 53% gave positive responses, while 20% gave negative responses (Figure 14).

**Conclusion #4:**

GEO-SEED would be a helpful system to support the geoscientist in their research and would be a helpful resource for students.
Overall Conclusion

Given the small sample size of 17, a 40% “Strongly Agree” positive response supports our conclusion that GEO-SEED would be a usable tool to support the geoscientist in sharing and discovering their Scientific Computational Entities. The 33% “Somewhat Agree” further supports this conclusion.

5.3 Performance Study

The experimental study reported in this section addresses the general question of whether GEO-SEED can handle large datasets. In particular, the study has two main goals: 1) to test the repository’s capability to answer queries for a dataset with 10,000 Web services annotations, and 2) to compare the query performance of the profile tables approach with that of the single table and property tables approaches.
Towards these goals, schema mapping and data mapping algorithms (e.g., see (Chebotko, 2007) for similar algorithms) were implemented in C++ to support the three approaches to repository design. RDF data was stored in a MySQL 5.1 CE RDBMS. The query translation (Chebotko, 2009) that was implemented in C++ was plugged in to support query mapping for Triple, property and profile tables.

The dataset for the experiments was used to generate 10,000 token “Web services.” Each Web service was on average annotated with 100 triples, resulting in overall 1,000,000 triples in the database.

The experiments were conducted on a PC with 3.00 GHz Intel Core 2 CPU, 4 GB RAM, and 750 GB disk space running MS Windows XP Professional.

The following section lists the sample SPARQL queries used in the performance experiment, followed by a sample dataset used as input. The corresponding SQL queries and additional sample datasets are included in APPENDIX E.

**Queries Used in the Study**

The five selected SPARQL queries, which are common for the GEO-SEED environment follow:

Q1 Find Web services that implement a computational entity with the name “gridding”.

Q2 Find Web services, along with their user reviews and overall quality-of-service ratings that implement a computational entity with the name “gridding”.

63
Q3 Find Web services that implement a computational entity with the name “gridding” and that have trust ≥ 4 and availability ratings ≥ 0.8.

Q4 Retrieve a general profile of a particular Web service.

Q5 Retrieve a quality-of-service profile of a particular Web service.

Q6 Retrieve quality-of-service profiles of two Web services for comparison.

**SPARQL Queries for Q1 through Q6.**

Corresponding SQL queries are presented in APPENDIX E.

**Q1** Find Web services that implement a computational entity with the name “gridding”.

PREFIX geoseed: <http://geoseed.cs.panam.edu/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?service
WHERE { ?process geoseed:name "gridding".
  ?service rdf:type geoseed:WebService .}

**Q2** Find Web services, along with their user reviews and overall ratings that implement a computational entity with the name “gridding”.

PREFIX geoseed: <http://geoseed.cs.panam.edu/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?service ?review ?rating
WHERE { ?process geoseed:name "gridding".
  ?service geoseed:describedBy ?profile .
  ?process rdf:type geoseed:ScientificComputationalEntity .}
Q3 Find Web services that implement a computational entity with the name "gridding" and that have trust >= 4 and availability ratings >= 0.8.

Q3 SPARQL
PREFIX geoseed: <http://geoseed.cs.panam.edu/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?service ?trust ?availability
WHERE { ?process geoseed:name "gridding".
  ?service geoseed:describedBy ?profile
  ?profile geoseed:trust ?trust
  ?profile rdf:type geoseed:QoSProfile.
  ?profile geoseed:availability ?availability
  FILTER (?trust >= 4 && ?availability >= 0.8) }

Q4 Retrieve a general profile of a particular Web service.

Q4 SPARQL

PREFIX geoseed: <http://geoseed.cs.panam.edu/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?profile ?pre ?obj
WHERE { geoseed:WS1 geoseed:describedBy ?profile .
  ?profile rdf:type geoseed:GeneralProfile .
  ?profile ?pre ?obj . }

Q5 Retrieve a quality-of-service profile of a particular Web service.

Q5 SPARQL

PREFIX geoseed: <http://geoseed.cs.panam.edu/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?profile ?pre ?obj
WHERE { geoseed:WS1 geoseed:describedBy ?profile .
    ?profile rdf:type geoseed:QoSProfile .
    ?profile ?pre ?obj . }

Q6 Retrieve quality-of-service profiles of two Web services for comparison.

Q6 SPARQL

PREFIX geoseed: <http://geoseed.cs.panam.edu/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?profile ?pre ?obj
WHERE { {{geoseed:WS1 geoseed:describedBy ?profile .}
    UNION
    {geoseed:WS5 geoseed:describedBy ?profile .}}
    ?profile rdf:type geoseed:QoSProfile .
    ?profile ?pre ?obj . }

Sample dataset used in the Performance Study. More are listed in APPENDIX E.

A sample data set used for the performance study follows. Variations were generated to be used as input into the performance simulation. The data set was based on the GeneralProfile and the QoSProfile

Sample 1 Data set for Performance Study

<SP1> <rdf:type> <ScientificComputationalEntity> .
<SP1> <name> "gridding" .
<WS1> <rdf:type> <WebService> .
<GP1> <rdf:type> <GeneralProfile> .
<QoSP1> <rdf:type> <QoSProfile> .
<WS1> <describedBy> <GP1> .
<WS1> <describedBy> <QoSP1> .
<SP1> <implementedBy> <WS1> .
<GP1> <name> "name of a web service" .
<GP1> <author> "author1" .
<GP1> <contact> "contact1" .
<GP1> <contributor> "contributor1" .
<GP1> <description> "description1" .
5.4 Study Results

Figure 15 reports the performance of the six test queries over the generated dataset, measured in microseconds, and stored using the three approaches: single table, property tables, and profile tables. The comparative results are also shown in the chart shown in Table 15. For the first three queries, the single table approach showed to be much slower than the other two approaches, since joins were performed on larger tables. On the other hand, property tables showed to be inefficient for queries Q4–Q6, where all profile descriptors were retrieved, because each property table that corresponded to a profile descriptor was accessed. Profile tables, which are a novel addition to our RDF repository, showed to be quite effective to handle all the six queries and especially queries Q4, Q5, and Q6.
Table 15: Time to Compute Queries

<table>
<thead>
<tr>
<th>Time to Complete Queries</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
</tr>
</thead>
<tbody>
<tr>
<td>single table</td>
<td>0.270667</td>
<td>0.828</td>
<td>0.609</td>
<td>0.213667</td>
<td>0.218667</td>
<td>0.218667</td>
</tr>
<tr>
<td>property tables</td>
<td>0.071333</td>
<td>0.265667</td>
<td>0.178333</td>
<td>0.75</td>
<td>0.666667</td>
<td>0.729333</td>
</tr>
<tr>
<td>profile tables</td>
<td>0.071333</td>
<td>0.25</td>
<td>0.16</td>
<td>0.124</td>
<td>0.124</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Figure 15: RDF Repository Query Performance

Conclusion

In summary, the performance study showed that GEO-SEED is capable to efficiently query large metadata collections. It also confirmed the usefulness of profile tables to substantially improve query performance.
CHAPTER 6
CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

The GEO-SEED architecture supports a new generation Web portal that can serve as a metadata repository for scientific computational entities in geosciences that is accessible from the Web and supports the scientist both for sharing and discovery and for machines for automating scientific workflow creation.

GEO-SEED will allow scientists to share their known applications and services by attaching metadata annotations and be able to discover relevant Web services used by other scientists in their field in a centralized comprehensive repository. Both scientist-friendly interface and machine-interpretable metadata are supported by the ontology-driven approach that organizes Web computational entity descriptors into six profiles. The dissertation presents the architecture of GEO-SEED and elaborates on the design and implementation of the structured wiki portal and the relational RDF repository of Web services annotations. It describes usability studies and presents a descriptive analysis and assessment of the GEO-SEED wiki usability. The usability study supported the stated objective to create a system that was easy for a geoscientist to use and provide support for the sharing and discovery of geosciences computational entities. It presents a unique profile table for storing the RDF data, based on horizontal partitioning. A performance study of the RDF repository was conducted. The performance study supports our conclusion that the GEO-SEED wiki provides a facility that researchers can use to efficiently
query large metadata collections and eventually use the results to create automated scientific workflows.

6.2 Future Work

In the future, our main focus will be on exploring user guided metadata extraction algorithms for the wiki. We will explore coupling GEO-SEED with an existing SWFMS, and we will extend the project to support annotation and discovery of scientific workflows and datasets in geosciences. Finally, we will refine the prototype to address user interface issues.
REFERENCES


78


APPENDIX A

SCREEN SHOT OF COMPLETED GEO-SEED ENTRY

The following pages show screen shots of the completed GEO-SEED entry for a sample Scientific Computational Entity, “The Generic Mapping Tool.”

Figure 16: Screen Shot Sample GEO-SEED Entry
Figure 17: Screen Shot GEO-SEED Entry – General Information

Figure 18: Screen Shot GEO-SEED Entry - Invocation and Quality of Service Information
Figure 19: Screen Shot GEO-SEED Entry - Deployment, Implementation, and Domain Knowledge Information
APPENDIX B
SURVEY OF GEOSCIENCE RESEARCHERS

This is a copy of the survey of Research Universities to gather information about how geology researchers currently share and discover geosciences computational entities they use in their research. Requests for information was sent to 10 institutions with Ph. D. programs in Geology, randomly selecting a mixture of faculty rank and a mixture of geographically placed institutions. Five institutions responded to the survey. The results are shown below.

GeoScience Research Survey
1. What is your research area?
   Earth Sciences, -
   Quaternary Geology -
   Planetary science -
   geoscience -
   geophysics -

2. What is your faculty rank?
   Professor 60.00% 3
   Associate Professor 0.00% 0
   Assistant Professor 40.00% 2
   Other (please specify)

3. How do you currently discover the computational entities you need for your research?
   Professional contacts 75.00% 3
   Online search engines 50.00% 2
   Literature Search 25.00% 1
   Geological Web Portals, List below.
   1. all
   Other

4. How do you or your colleagues currently share computational entities you use or have created?
   Professional Publication 40.00% 2
   Personal Web Pages 40.00% 2
   Professional Contacts 20.00% 1
Have not created computational entities 20.00% 1
Other (please specify) 40.00% 2

1. first 3
2. software repositories

5. Are you aware of any existing comprehensive Web support system that enables geoscientists to share and discover computational entities?

Not aware of any 60.00% 3
If aware, please list 40.00% 2
http://www.orfeus-eu.org/
http://www.iris.edu/software/
geodynamics.org
APPENDIX C
GEO-SEED USABILITY STUDY I

The following pages show the documents we handed out for the GEO-SEED User Case Study I. This was a paper and pencil exercise to give preliminary feedback on the concept of the Discovery Ontology approach for sharing information. We asked for Background information of the participants. Included below are copies of the forms they would fill in for the General Information, Deployment Information, and QoS Information, which include entries for the various descriptors for the web service they were registering. We also asked the participants to indicate if they would need help explaining the meaning of the descriptors for each category.

We gave a scenario for a geologist who was interested in creating a contour map using a gridding process. Assuming the geologist would search the web for available software and then want to share what he or she found by registering it in GEO-SEED, we gave paper copies of such a Website, which contained Geology Contour Mapping Information and Directions on how to download software from the site.

Background Information Form

<table>
<thead>
<tr>
<th>Participant ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profile: (student, major,[fr,so,jr,sr]), other (Specify)</td>
</tr>
<tr>
<td>Internet usage: Select all that apply</td>
</tr>
<tr>
<td>a. Look up definitions</td>
</tr>
<tr>
<td>b. Find software</td>
</tr>
</tbody>
</table>
c. Contribute to blogs, etc (Describe)  
d. Find procedure to solve a specific problem  
e. Other (Describe)  

Extent of your Knowledge of Geology Field

On a scale of 1..5 rate your knowledge of  
a. Web application  
b. Web Service  
c. Work flows

Scenario:

You are a geologist and you want to create a contour map showing the rock formation under the earth in the South West region of the United States. You have a work flow for the task that shows that you use gravity data from a data file. This data must be corrected with Bouguer Anomaly process. Once that is done the data is sent to a gridding process with the results sent to the contouring process to create the contour map. You want to find an algorithm or program. You Google “contour maps” and get a description as follows:

You were advised to use GMT gridding. Google “gmt gridding” and you find the following site:

http://gmt.soest.hawaii.edu/

When you select download from the left menu, you get the screen with deployment information.

See attached Web pages:
After clicking the Download button above you get the following screen:

### Obtaining and Installing GMT

GMT is available via anonymous ftp from a global set of ftp servers; each contain the same files as the main server in Hawaii. File transfer is usually faster if you select the server closest to you. Our installer will automatically get the required archives from the ftp site you choose.

Users with a slow Internet-connection and users who desire large amounts of supplemental data sets ready to be used with GMT: See the GMT Companion DVD-R products distributed by Geoware.
1. **Fast-track for (repeat) UNIX/Linux/OSX users**

I've done this before. Take me directly to the *INSTALL FORM*.

---

2. **Platform-specific Instructions**

3. **UNIX or LINUX**: Note: The install process requires *bzip2*.

   1. *Automated install [Recommended]*. Obtain and install GMT by interacting with the *INSTALL FORM*. Follow instructions there to obtain the Bourne shell install-script and a customized install parameter file. The automated install will also install netCDF if needed.

   2. *Manual install*. If you prefer, you can also do the typical manual install by ftp'ing the files, untar, run configure, make etc. Read the README file for the required steps. For manual install you must also manually get and install the *Unidata netCDF library* which GMT requires, or have the library already installed. Use ftp to any of the GMT *mirror sites*.

   3. *CVS installation for GMT gurus*. To get the bleeding edge GMT version and even contribute to the development of GMT, consider installing the "live" GMT version by following the *CVS Instructions*. 
1. **WINDOWS:**

1. *DOS batch files rule.* If you just want to install Windows executables and get on with it, visit our GMT Windows page for access to Windows Installers. Note that many of the DOS example scripts utilize GNU awk; the WIN32 executable *gawk* has therefore been placed on all ftp sites.

2. *DOS batch files suck, part I.* Because you cannot get much done with DOS batch jobs, we strongly recommend that you install Cygwin, a free UNIX emulation package for Windows. Cygwin lets you open shell windows and access standard UNIX tools such as tcsh, gcc, etc. You would then install GMT as described above for UNIX/Linux.

3. *DOS batch files suck, part II.* If you run Windows, you can get access to csh command windows by installing the freely available Windows Services for UNIX, a UNIX environment for Windows. SFU lets you install GMT as described for UNIX/Linux above.

4. *DOS batch files suck, part III.* Finally, you may consider the option of purchasing and installing *VMWare*, a virtual computer which allows you to install numerous operating systems (such as Linux) and then pursue the general Linux/UNIX install option above within your virtual machine.

2. **OS X:** GMT installs and runs under Apple's OS X which is UNIX-based; just follow instructions for UNIX/Linux above. You must first install the Xcode Developer Tools (which includes the GNU C compiler, make, etc) as these are not installed by default but is an optional install via the OS X Install DVD. You can also download them from Apple's support site. Also select to install X11. Finally, the latest versions of GMT are also available as user-friendly packages via *Fink*.

3. **OS/2:** GMT has been ported to OS/2. For information and precompiled executables, see Allen Cogbill's GMT OS/2 page.
From the information on how to Use GEO-SEED, you find you need to create a General Information page and Deployment page and add comments to the QoS page. Fill out the paper forms to the best of your ability with what you have found for each descriptor for the Gridding Software you have found.

4. **GEO-SEED Forms**

<table>
<thead>
<tr>
<th>General Information</th>
<th>Need Help to explain field</th>
<th>Enter Values Here</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Y / N</td>
<td></td>
</tr>
<tr>
<td>Authors</td>
<td>Y / N</td>
<td></td>
</tr>
<tr>
<td>Contact Information</td>
<td>Y / N</td>
<td></td>
</tr>
<tr>
<td>Contributors</td>
<td>Y / N</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>Y / N</td>
<td></td>
</tr>
<tr>
<td>URLs</td>
<td>Y / N</td>
<td></td>
</tr>
<tr>
<td>Unique Identifier</td>
<td>Y / N</td>
<td></td>
</tr>
<tr>
<td>Version</td>
<td>Y / N</td>
<td></td>
</tr>
<tr>
<td>Release Date</td>
<td>Y / N</td>
<td></td>
</tr>
<tr>
<td>Languages</td>
<td>Y / N</td>
<td></td>
</tr>
<tr>
<td>Subject</td>
<td>Y / N</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>Y / N</td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td>Y / N</td>
<td></td>
</tr>
<tr>
<td>License</td>
<td>Y / N</td>
<td></td>
</tr>
<tr>
<td>Support</td>
<td>Y / N</td>
<td></td>
</tr>
</tbody>
</table>
### QoS Information

<table>
<thead>
<tr>
<th>Field</th>
<th>Y / N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trust</td>
<td></td>
</tr>
<tr>
<td>Reliability</td>
<td></td>
</tr>
<tr>
<td>Availability</td>
<td></td>
</tr>
<tr>
<td>Processing Time</td>
<td></td>
</tr>
<tr>
<td>Requests per Second</td>
<td></td>
</tr>
<tr>
<td>Custom Metric</td>
<td></td>
</tr>
<tr>
<td>Security</td>
<td></td>
</tr>
<tr>
<td>Known Failures</td>
<td></td>
</tr>
<tr>
<td>Overall Rating: 1..5 (top)</td>
<td></td>
</tr>
<tr>
<td>User Reviews (comments)</td>
<td></td>
</tr>
</tbody>
</table>

### Deployment Information

<table>
<thead>
<tr>
<th>Field</th>
<th>Y / N</th>
</tr>
</thead>
<tbody>
<tr>
<td>URLs</td>
<td></td>
</tr>
<tr>
<td>Software Architecture</td>
<td></td>
</tr>
<tr>
<td>Operating Systems</td>
<td></td>
</tr>
<tr>
<td>Software Dependencies</td>
<td></td>
</tr>
<tr>
<td>Hardware</td>
<td></td>
</tr>
<tr>
<td>Installation</td>
<td></td>
</tr>
<tr>
<td>Activation and Deactivation</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX D
USABILITY STUDY II

D.1 GEO-SEED Usability Study

By Pearl Brazier
Contact: brazier@utpa.edu 956-381-3455

Introduction: I am completing my Ph. D. in Computer Science with Dr. Ann Gates at UTEP. As part of my research, I have developed a Wiki prototype system, GEO-SEED, to assist Geoscientists to share and discover software applications and services. As part of that endeavor I am asking Geoscientists to evaluate the prototype for its usability. The evaluation should take no longer than one hour.

Steps to complete the Usability Study:
1. Access geoseed.cs.panam.edu using your browser.
   http://geoseed.cs.panam.edu
2. Register as a new user. Remember your username and password.
3. Login in.
4. Submit a contribution to GEO-SEED, filling out all the information you have available.
   1) a. If you currently use an application or have developed one, register it.
      The type may be application, webservice, dataset, web portal or other [Specify].
   Or
   1) b. You may search with your web browser to find geological related candidates.
      Suggested areas may be map making, seismic data, rock formations, geology portals, remote-sensing, web mapping tools.
   2) Register what you found.
      Steps to register a contribution:
      After logging in:
      1) Select Fill out the GEO-SEED form by clicking on the Click Here link on the home page.
      2) Fill out the form as completely as you can for the information that was available.
      3) Select the Save and View button.
      4) Logout.
      5) Log back in and search for the contribution you just entered.
5. Search for Talwani and view the information.
6. Search for GEON or GEONgrid and view the information. Then access the URL that is listed for GEON.

   Add your rating for GEON or the entry you just registered under the Quality of Service section.

7. Now rate your usability experience by completing the survey at the following:
   http://www.surveymonkey.com/s/W53JGM8

8. Thank you for participating.
## D.2 Usability Survey Results

By Pearl Brazier

July 13, 2010

<table>
<thead>
<tr>
<th></th>
<th>sent</th>
<th>returned</th>
<th>return rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey Response Rate</td>
<td>31</td>
<td>17</td>
<td>55%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Professional Status</th>
<th>Researcher</th>
<th>Professor</th>
<th>Ph. D. Student</th>
<th>Master's Student</th>
<th>Undergraduate Student</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of Whole (count)</td>
<td>6.7% (1)</td>
<td>26.7% (4)</td>
<td>6.7% (1)</td>
<td>13.3% (2)</td>
<td>40.0% (6)</td>
<td>6.7% (1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Education Level</th>
<th>Ph. D. Student</th>
<th>Master's Student</th>
<th>Undergraduate Student</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of Whole (count)</td>
<td>37.5% (6)</td>
<td>18.8% (3)</td>
<td>37.5% (6)</td>
<td>6.3% (1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Computer/Internet Experience</th>
<th>Experienced Technical User</th>
<th>Experienced General User</th>
<th>Little Experience</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of Whole (Count)</td>
<td>66.7% (10)</td>
<td>33.3% (5)</td>
<td>0.0% (0)</td>
<td>0.0% (0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Geology Knowledge</th>
<th>No Knowledge</th>
<th>Little Knowledge</th>
<th>Knowledge in one area</th>
<th>Knowledge in several areas</th>
<th>Extensive Knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of Whole (Count)</td>
<td>14.3% (2)</td>
<td>42.9% (6)</td>
<td>14.3% (2)</td>
<td>0.0% (0)</td>
<td>28.6% (4)</td>
</tr>
<tr>
<td>--------------------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>Quaternary Geology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tomography</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>geophysics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>geophysics - gravity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>not much</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computer Science</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structural Geology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computer Science</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic Web</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mineral chemistry</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mineral chemistry</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>% of Whole (count)</th>
<th>yes</th>
<th>86.70%</th>
<th>13</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>13.30%</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>not sure</td>
<td>0.00%</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I have used a Wiki to find information</td>
<td>yes</td>
<td>85.7% (12)</td>
<td>14.3% (2)</td>
<td>0.0% (0)</td>
<td>14</td>
</tr>
<tr>
<td>Question</td>
<td>Yes (%)</td>
<td>No (%)</td>
<td>Not Sure (%)</td>
<td>Response Count</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>---------</td>
<td>--------</td>
<td>--------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td>I understand the difference between information found in a Wiki and information found on a Web page</td>
<td>78.6% (11)</td>
<td>14.3% (2)</td>
<td>7.1% (1)</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>I have found information on a Web Portal</td>
<td>71.4% (10)</td>
<td>7.1% (1)</td>
<td>21.4% (3)</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>I usually find information online using a Search Engine</td>
<td>92.9% (13)</td>
<td>0.0% (0)</td>
<td>7.1% (1)</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Search Engine</td>
<td>Google (9)</td>
<td>Yahoo (2)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes (%)</th>
<th>No (%)</th>
<th>Not Sure (%)</th>
<th>Response Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>I know what a Web service is</td>
<td>85.7% (12)</td>
<td>14.3% (2)</td>
<td>0.0% (0)</td>
<td>14</td>
</tr>
<tr>
<td>I have heard about WSDL</td>
<td>78.6% (11)</td>
<td>14.3% (2)</td>
<td>7.1% (1)</td>
<td>14</td>
</tr>
<tr>
<td>I know what a workflow is</td>
<td>71.4% (10)</td>
<td>7.1% (1)</td>
<td>21.4% (3)</td>
<td>14</td>
</tr>
<tr>
<td>I have heard about Web 2.0 Technologies</td>
<td>92.9% (13)</td>
<td>0.0% (0)</td>
<td>7.1% (1)</td>
<td>14</td>
</tr>
</tbody>
</table>
7. Please rate each of the following tasks in registering information in GEO-SEED relative to their ease of use.

<table>
<thead>
<tr>
<th>Task</th>
<th>Strongly Disagree</th>
<th>Somewhat Disagree</th>
<th>No Opinion</th>
<th>Somewhat Agree</th>
<th>Strongly Agree</th>
<th>N/A</th>
<th>Rating Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>I experienced no difficulty registering as a GEO-SEED user</td>
<td>20.0% (3)</td>
<td>0.0% (0)</td>
<td>0.0% (0)</td>
<td>13.3% (2)</td>
<td>66.7% (10)</td>
<td>0.0% (0)</td>
<td>4.07</td>
</tr>
<tr>
<td>I experienced no difficulty logging in to GEO-SEED</td>
<td>6.3% (1)</td>
<td>18.8% (3)</td>
<td>0.0% (0)</td>
<td>6.3% (1)</td>
<td>68.8% (11)</td>
<td>0.0% (0)</td>
<td>4.13</td>
</tr>
<tr>
<td>It was clear how to get started to register my information in GEO-SEED</td>
<td>33.3% (5)</td>
<td>0.0% (0)</td>
<td>13.3% (2)</td>
<td>33.3% (5)</td>
<td>20.0% (3)</td>
<td>0.0% (0)</td>
<td>3.07</td>
</tr>
<tr>
<td>I was able to add to or change existing information in GEO-SEED</td>
<td>7.1% (1)</td>
<td>28.6% (4)</td>
<td>7.1% (1)</td>
<td>14.3% (2)</td>
<td>28.6% (4)</td>
<td>14.3% (2)</td>
<td>3.33</td>
</tr>
<tr>
<td>It was easy to learn to use the system</td>
<td>7.1% (1)</td>
<td>35.7% (5)</td>
<td>7.1% (1)</td>
<td>21.4% (3)</td>
<td>28.6% (4)</td>
<td>0.0% (0)</td>
<td>3.29</td>
</tr>
<tr>
<td>The information provided how to use the system was adequate</td>
<td>30.8% (4)</td>
<td>15.4% (2)</td>
<td>7.7% (1)</td>
<td>30.8% (4)</td>
<td>15.4% (2)</td>
<td>0.0% (0)</td>
<td>2.85</td>
</tr>
</tbody>
</table>
8. Rate the following as to the degree of difficulty you encountered completing the form to register new Information in GEO-SEED.

<table>
<thead>
<tr>
<th>Information Type</th>
<th>No Difficulty</th>
<th>Some Difficulty</th>
<th>Extreme Difficulty</th>
<th>Most Information Available</th>
<th>Some Information Available</th>
<th>No Information Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Information</td>
<td>57.1% (8)</td>
<td>28.6% (4)</td>
<td>7.1% (1)</td>
<td>14.3% (2)</td>
<td>21.4% (3)</td>
<td>0.0% (0)</td>
</tr>
<tr>
<td>Invocation Information</td>
<td>50.0% (7)</td>
<td>35.7% (5)</td>
<td>0.0% (0)</td>
<td>7.1% (1)</td>
<td>21.4% (3)</td>
<td>7.1% (1)</td>
</tr>
<tr>
<td>Deployment Information</td>
<td>57.1% (8)</td>
<td>21.4% (3)</td>
<td>7.1% (1)</td>
<td>14.3% (2)</td>
<td>14.3% (2)</td>
<td>7.1% (1)</td>
</tr>
<tr>
<td>Quality of Service</td>
<td>42.9% (6)</td>
<td>35.7% (5)</td>
<td>7.1% (1)</td>
<td>7.1% (1)</td>
<td>21.4% (3)</td>
<td>7.1% (1)</td>
</tr>
<tr>
<td>Implementation Information</td>
<td>42.9% (6)</td>
<td>35.7% (5)</td>
<td>7.1% (1)</td>
<td>7.1% (1)</td>
<td>21.4% (3)</td>
<td>7.1% (1)</td>
</tr>
<tr>
<td>Domain Knowledge Information</td>
<td>50.0% (7)</td>
<td>35.7% (5)</td>
<td>7.1% (1)</td>
<td>7.1% (1)</td>
<td>28.6% (4)</td>
<td>0.0% (0)</td>
</tr>
</tbody>
</table>
- For Quality of Service information, it may be difficult for a "geologist" to come up with this information. It may even be difficult for a technical person. For example, if the web service being registered is a prototype intended to explore a research idea. Is your site intended for registration of well established web services only? If so, this needs to be stated somewhere. Furthermore, the Quality of Service information may be very dynamic and susceptible to a wide variety of factors, such as network traffic and demand levels. Also, what do you mean by "trust"? And what is the scale for "overall rating? Are there specific guidelines that you are expecting me to use to "rate" my service consistently?

- Deployment information: What information is expected in "activation and deactivation"? Does hardware refer to the platform from which I am calling the service or the platform from which I am running the service? What links are you expecting in the "URLs" field? What are the options for "Software Architecture"? I am not sure how Software Architecture fits into the "Deployment" section.

9. Please rate the following features of GEO-SEED Wiki

<table>
<thead>
<tr>
<th></th>
<th>Strongly Disagree</th>
<th>Somewhat Disagree</th>
<th>Neither Agree or Disagree</th>
<th>Some Agree</th>
<th>Strongly Agree</th>
<th>N/A</th>
<th>Rating Average</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>I successfully searched for information I was interested in</td>
<td>6.3% (1)</td>
<td>6.3% (1)</td>
<td>0.0% (0)</td>
<td>25.0% (4)</td>
<td>56.3% (9)</td>
<td>6.3% (1)</td>
<td>4.27</td>
<td>16</td>
</tr>
<tr>
<td>I was able to understand the fields in each of the categories</td>
<td>14.3% (2)</td>
<td>21.4% (3)</td>
<td>0.0% (0)</td>
<td>35.7% (5)</td>
<td>28.6% (4)</td>
<td>0.0% (0)</td>
<td>3.43</td>
<td>14</td>
</tr>
<tr>
<td><strong>I was able to find help when I needed it</strong></td>
<td>12.5% (2)</td>
<td>12.5% (2)</td>
<td>18.8% (3)</td>
<td>12.5% (2)</td>
<td>25.0% (4)</td>
<td>18.8% (3)</td>
<td>3.31</td>
<td>16</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>------</td>
<td>-----</td>
</tr>
<tr>
<td>GEO-SEED would be helpful for sharing processes a Geologist uses</td>
<td>6.7% (1)</td>
<td>13.3% (2)</td>
<td>6.7% (1)</td>
<td>26.7% (4)</td>
<td>46.7% (7)</td>
<td>0.0% (0)</td>
<td>3.93</td>
<td>15</td>
</tr>
<tr>
<td>GEO-SEED would be helpful for sharing existing datasets information</td>
<td>13.3% (2)</td>
<td>6.7% (1)</td>
<td>6.7% (1)</td>
<td>20.0% (3)</td>
<td>53.3% (8)</td>
<td>0.0% (0)</td>
<td>3.93</td>
<td>15</td>
</tr>
<tr>
<td>GEO-SEED would be helpful for sharing geology application software information</td>
<td>13.3% (2)</td>
<td>6.7% (1)</td>
<td>6.7% (1)</td>
<td>20.0% (3)</td>
<td>53.3% (8)</td>
<td>0.0% (0)</td>
<td>3.93</td>
<td>15</td>
</tr>
<tr>
<td>GEO-SEED would be helpful for locating web portals a Geologist might use</td>
<td>13.3% (2)</td>
<td>6.7% (1)</td>
<td>6.7% (1)</td>
<td>20.0% (3)</td>
<td>53.3% (8)</td>
<td>0.0% (0)</td>
<td>3.93</td>
<td>15</td>
</tr>
</tbody>
</table>
**GEO-SEED would be helpful for students studying Geology**

<table>
<thead>
<tr>
<th>Overall GEO-SEED would be a useful tool for sharing and searching for Geological Information</th>
<th>20.0% (3)</th>
<th>6.7% (1)</th>
<th>0.0% (0)</th>
<th>33.3% (5)</th>
<th>40.0% (6)</th>
<th>3.67</th>
<th>15</th>
</tr>
</thead>
</table>

10. Please give an overall rating for the GEO-SEED Wiki. List the most negative and the most positive aspects in the Comment box. Give any general comments about GEO-SEED

<table>
<thead>
<tr>
<th>answered question</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>skipped question</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Somewhat Disagree</th>
<th>Neither Agree or Disagree</th>
<th>Somewhat Agree</th>
<th>Strongly Agree</th>
<th>Rating Response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Average Count</td>
</tr>
</tbody>
</table>

| Overall GEO-SEED would be a useful tool for sharing and searching for Geological Information | 20.0% (3) | 6.7% (1) | 0.0% (0) | 33.3% (5) | 40.0% (6) | 3.67 | 15 |
D.3 Descriptive Statistics for Usability Survey

Methods used for analysis

While the data set is small, a descriptive analysis can be done, which presents conclusions based on the percentages of responses in the categories for each question. Two statistical methods of analysis were used. In the first method, the categories were grouped into two categories: Category 1: Agree and Strongly Agree and category 2: Strongly Disagree and Disagree, ignoring the No Opinion responses. While technically not a binomial distribution, we assumed that it was equally likely for a participant to agree or disagree, therefore we present the results of a binomial distribution test on the data. As part of a binomial distribution, we calculated the p value for questions 7 through 10. It is standard practice to consider a p value of less than 0.05 as significant in a binomial distribution. The resulting p values are shown below. For the null hypothesis that an equal number of users consider GEO-SEED to be a usable system for the geosciences community, we will reject the null hypothesis for p values less than or equal to 0.05. To assess results, we marked each question as SUPPORT or NOT SUPPORT our conclusions that the number of users who support this conclusion is statistically significantly larger than the number of those who do not.

We also conducted a t test on the data where we considered the four categories of Strongly Disagree, Disagree, Agree, and Strongly Agree so that the analysis reflects the range of responses. We represented the data respectively as -2, -1, 1, and 2, with a degree of freedom one less than the total number of responses. The p values were similar to the
results we found using the binomial distribution test, resulting in no changes to our conclusions.

In particular, for the overall rating results for the Binomial Distribution and the t test calculation for Question 10, in which participants were asked to give an overall rating for the usability of GEO-SEED, this question showed a BINOM-DIST p value of 0.059 and for the t test / 0.06, which was sufficiently close to 0.050 to support our conclusion that GEO-SEED would be a useful system. Note: if we used a sample size of 17 reflecting the fact that not all participants answered this question, the p value is 0.025. The binomial distribution and t test p values for question 9, which asked for feedback on specific uses of GEO-SEED all had p-values ranging from 0.02 to 0.046, further supporting our conclusion.

**Descriptive analysis of GEO-SEED detailed usability questions**

In the other more detailed usability questions 7 through 9, we conclude that statistically significant more users responded positively than negatively that registering and logging into GEO-SEED was easy, while getting started submitting entities, changing entities, learning how to use the system, and having adequate information on how to use the system responses result in such a significant majority (Figure 11). The response to the question concerning the degree of difficulty for completing the descriptors for each profile is given in Figure 12. We conclude that we need to revise the help and user interface features that are presently implemented in GEO-SEED. The response to question 9 to rate the features of GEO-SEED, except for the category of understanding the fields of each category, strongly supported our conclusion that GEO-SEED would be a useful system for the sharing geosciences computational entities (Figure 13 and Figure 14).
Table showing the results of binomial distribution p-values and t-test p-values for various tasks in registering information in GEO-SEED.

<table>
<thead>
<tr>
<th>Question Part</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>No Opinion</th>
<th>Agree</th>
<th>Strongly Agree</th>
<th>Total</th>
<th>BINOM DIST p value</th>
<th>t test p value</th>
<th>t test mean</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>I experienced no difficulty registering as a GEO-SEED user</td>
<td>3</td>
<td>2</td>
<td>10</td>
<td>15</td>
<td></td>
<td>0.02</td>
<td>0.02</td>
<td>1.07</td>
<td>Support</td>
<td></td>
</tr>
<tr>
<td>I experienced no difficulty logging in to GEO-SEED</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>16</td>
<td></td>
<td>0.04</td>
<td>0.004</td>
<td>1.13</td>
<td>Support</td>
<td></td>
</tr>
<tr>
<td>It was clear how to get started to register[submit] my information in GEO-SEED</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td></td>
<td>0.15</td>
<td>0.44</td>
<td>0.08</td>
<td>Not Support</td>
<td></td>
</tr>
<tr>
<td>I was able to add to or change existing information in GEO-SEED</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td></td>
<td>0.39</td>
<td>0.23</td>
<td>0.36</td>
<td>Not Support</td>
<td></td>
</tr>
<tr>
<td>It was easy to learn to use the system</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td></td>
<td>0.40</td>
<td>0.24</td>
<td>0.31</td>
<td>Not Support</td>
<td></td>
</tr>
<tr>
<td>The information provided how to use the system was adequate</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td></td>
<td>0.50</td>
<td>0.63</td>
<td>0.366</td>
<td>Not Support</td>
<td></td>
</tr>
</tbody>
</table>

For t test, data coding scheme: Strongly Disagree -2, Disagree -1, Agree 1, Strongly Agree 2
The No Opinion category was not counted.
8. Rate the following as to the degree of difficulty you encountered completing the form to register new information in GEO-SEED.

<table>
<thead>
<tr>
<th>Question</th>
<th>Part</th>
<th>No Difficulty</th>
<th>Some Difficulty</th>
<th>Extreme Difficulty</th>
<th>Total</th>
<th>BINOM DIST p value</th>
<th>t test p value</th>
<th>t test mean</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Information</td>
<td>8 A</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>13</td>
<td>0.29</td>
<td>0.21</td>
<td>0.23</td>
<td>Not Support</td>
</tr>
<tr>
<td>Invocation Information</td>
<td>8 B</td>
<td>7</td>
<td>5</td>
<td>0</td>
<td>12</td>
<td>0.39</td>
<td>0.29</td>
<td>0.17</td>
<td>Not Support</td>
</tr>
<tr>
<td>Deployment Information</td>
<td>8 C</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td>12</td>
<td>0.19</td>
<td>0.13</td>
<td>0.33</td>
<td>Not Support</td>
</tr>
<tr>
<td>Quality of Service Information</td>
<td>8 D</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>12</td>
<td>0.61</td>
<td>0.50</td>
<td>0.00</td>
<td>Not Support</td>
</tr>
<tr>
<td>Implementation Information</td>
<td>8 E</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>12</td>
<td>0.61</td>
<td>0.50</td>
<td>0.00</td>
<td>Not Support</td>
</tr>
<tr>
<td>Domain Knowledge Information</td>
<td>8 F</td>
<td>7</td>
<td>5</td>
<td>1</td>
<td>13</td>
<td>0.50</td>
<td>0.40</td>
<td>0.08</td>
<td>Not Support</td>
</tr>
<tr>
<td>Question</td>
<td>Part</td>
<td>Strongly Disagree</td>
<td>Disagree</td>
<td>No Opinion</td>
<td>Agree</td>
<td>Strongly Agree</td>
<td>Total</td>
<td>BINOM DIST</td>
<td>t test p value</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>------</td>
<td>-------------------</td>
<td>----------</td>
<td>------------</td>
<td>--------</td>
<td>----------------</td>
<td>-------</td>
<td>-------------</td>
<td>---------------</td>
</tr>
<tr>
<td>I successfully searched for information I was interested in</td>
<td>9 A</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>9</td>
<td>15</td>
<td>0.004</td>
<td>0.006</td>
</tr>
<tr>
<td>I was able to understand the fields in each of the categories</td>
<td>9 B</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>14</td>
<td>0.21</td>
<td>0.13</td>
</tr>
<tr>
<td>I was able to find help when I needed it</td>
<td>9 C</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>13</td>
<td>0.13</td>
<td>0.24</td>
</tr>
<tr>
<td>GEO-SEED would be helpful for sharing processes a Geologist uses</td>
<td>9 D</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>15</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>GEO-SEED would be helpful for sharing existing datasets information</td>
<td>9 E</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>8</td>
<td>15</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>GEO-SEED would be helpful for sharing geology application software information</td>
<td>9 F</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>8</td>
<td>15</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>GEO-SEED would be helpful for</td>
<td>9 G</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>13</td>
<td>0.05</td>
<td>0.05</td>
</tr>
</tbody>
</table>
This question asks the participants to give an overall rating for the usability of the GEO-SEED wiki.

<table>
<thead>
<tr>
<th>Question</th>
<th>Part</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>No Opinion</th>
<th>Agree</th>
<th>Strongly Agree</th>
<th>Total</th>
<th>BINOM</th>
<th>DIST</th>
<th>p value</th>
<th>t test p value</th>
<th>t test mean</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall GEO-SEED would be a useful tool for sharing and searching for Geological Information</td>
<td>10</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>6</td>
<td>15</td>
<td>0.059</td>
<td>0.06</td>
<td>0.67</td>
<td>Support</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX E
PERFORMANCE STUDY

SQL Query Equivalencies for SPARQL queries

Q1 Find Web services that implement a computational entity with the name “gridding”.

Q1 SPARQL

PREFIX geoseed: <http://geoseed.cs.panam.edu/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?service
WHERE { ?process geoseed:name "gridding".
  ?service rdf:type geoseed:WebService . }

Q1 TRIPLE

Select t1.o
From rdf_Triple t1, rdf_Triple t2
Where t1.p = 'implementedBy' and t2.p = 'name' and t1.s = t2.s and t2.o = 'gridding'

Q1 PROPERTY

Select t1.o
From rdf_implementedby t1, rdf_name t2
Where t1.s = t2.s and t2.o = 'gridding'

Q1 PROFILE

The same as Q1 PROPERTY because the query is not a profile-related query

=================================================================
Q2 Find Web services, along with their user reviews and overall quality-of-service ratings that implement a computational entity with the name “gridding”.

Q2 SPARQL

PREFIX geoseed: <http://geoseed.cs.panam.edu/>  
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>  
SELECT ?service ?review ?rating  
WHERE {  
  ?process geoseed:name "gridding" .  
  ?service geoseed:describedBy ?profile .  
  ?profile rdf:type geoseed:QoSProfile . }

Q2 TRIPLE

Select t1.o, t4.o, t5.o  
From rdf_Triple t1, rdf_Triple t2, rdf_Triple t3, rdf_Triple t4, rdf_Triple t5  
Where t1.p = 'implementedBy' and t2.p = 'name' and t1.s = t2.s and t2.o = 'gridding'  
and t3.p = 'describedby' and t1.o = t3.s and t4.p = 'userReview'  
and t3.o = t4.s  
and t5.p = 'overallRating' and t3.o = t5.s

Q2 PROPERTY

Select t1.o, t4.o, t5.o  
From rdf_implementedBy t1, rdf_name t2, rdf_describedby t3, rdf_userReview t4,  
rdf_overallRating t5  
Where t1.s = t2.s and t2.o = 'gridding' and t1.o = t3.s and t3.o = t4.s and  
t3.o = t5.s

Q2 PROFILE

Select t1.o, t4.o  
From rdf_implementedBy t1, rdf_name t2, rdf_describedby t3,
Q3 SPARQL

PREFIX geoseed: <http://geoseed.cs.panam.edu/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?service ?trust ?availability
WHERE {
  ?process geoseed:name "gridding".
  ?service geoseed:describedBy ?profile.
  ?profile rdf:type geoseed:QoSProfile.
  FILTER (?trust >= 4 && ?availability >= 0.84)
}

Q3 TRIPLE

Select t1.o, t4.o, t5.o
From rdf_implementedBy t1, rdf_name t2, rdf_describedBy t3, rdf_trust t4,
Where t1.p = 'implementedBy' and t2.p = 'name' and t1.s = t2.s and t2.o = 'gridding'
and t3.p = 'describedby' and t1.o = t3.s and t4.p = 'trust' and t3.o = t4.s
and t5.p = 'availability' and t3.o = t5.s and t4.o >= '4' and t5.o >= '0.8'

Q3 PROPERTY

Select t1.o, t4.o, t5.o
From rdf_implementedBy t1, rdf_name t2, rdf_describedby t3, rdf_trust t4,
rdf_availability t5
Where t1.s = t2.s and t2.o = 'gridding' and t1.o = t3.s and t3.o = t4.s and t3.o = t5.s and t4.o >= '4' and t5.o >= '0.8'

Q3 PROFILE

Select t1.o, t4.o, t5.o
From rdf_implementedBy t1, rdf_name t2, rdf_describedby t3, rdf_qosprofilessubject t4, rdf_qosprofilessubject t5
Where t1.s = t2.s and t2.o = 'gridding' and t1.o = t3.s and t3.o = t4.i and t4.p='trust' and t4.o >= '4' and t3.o = t5.i and t5.p='availability' and t5.o >= '0.8'

Q4 Retrieve a general profile of a particular Web service.

Q4 SPARQL

PREFIX geoseed: <http://geoseed.cs.panam.edu/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?profile ?pre ?obj
WHERE { geoseed:WS1 geoseed:describedBy ?profile .
  ?profile rdf:type geoseed:GeneralProfile .
  ?profile ?pre ?obj .}

Q4 TRIPLE

Select t3.s, t3.p, t3.o
From rdf_Triple t1, rdf_Triple t2, rdf_Triple t3
Where t1.s = 'WS1' and t1.p = 'describedBy' and t2.s = t1.o and t2.p='rdf:type' and t2.o = 'GeneralProfile' and t3.s = t2.s

Q4 PROPERTY

select x2.s,x2.o
from

(Select t2.s as s
From rdf_describedBy t1, rdf_type t2
Where t1.s = 'WS1' and t2.s = t1.o
and t2.o = 'GeneralProfile') x1

inner join {
select * from
rdf_type
union
select * from
rdf_name
union
select * from
rdf_author
union
select * from
rdf_contact
union
select * from
rdf_contributor
union
select * from
rdf_description
union
select * from
rdf_url
union
select * from
rdf_identifier
union
select * from
rdf_version
union
select * from
rdf_releaseDate
union

}
select * from rdf_language 
union
select * from rdf_cost 
union
select * from rdf_license 
union
select * from rdf_support 
) x2

on (x1.s = x2.s)

Q4 PROFILE

Select t2.i, t2.p, t2.o
From rdf_describedBy t1, rdf_generalprofilesobject t2
Where t1.s = 'WS1'
and t1.o = t2.i

=================================================================

Q5 Retrieve a quality-of-service profile of a particular Web service.

Q5 SPARQL

PREFIX geoseed: <http://geoseed.cs.panam.edu/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?profile ?pre ?obj
WHERE { geoseed:WS1 geoseed:describedBy ?profile .
   ?profile rdf:type geoseed:QoSProfile .
   ?profile ?pre ?obj .}

Q5 TRIPLE

Select t3.s, t3.p, t3.o
From rdf_Triple t1, rdf_Triple t2, rdf_Triple t3
Where t1.s = 'WS1' and t1.p = 'describedBy' and t2.s = t1.o and t2.p='rdf:type'
and t2.o = 'QoSProfile' and t3.s = t2.s

Q5 PROPERTY

select x2.s,x2.o
from
(Select t2.s as s
From rdf_describedBy t1, rdf_type t2
Where t1.s = 'WS1' and t2.s = t1.o
and t2.o = 'QoSProfile') x1

inner join {
select * from
rdf_type
union
select * from
rdf_trust
union
select * from
rdf_reliability
union
select * from
rdf_availability
union
select * from
rdf_processingTime
union
select * from
rdf_requestspersecond
union
select * from
rdf_security
union
select * from
rdf_failure
union
select * from

union
select * from rdf_overallrating
union
select * from rdf_userreview
) x2

on (x1.s = x2.s)

Q5 PROFILE

Select t2.i, t2.p, t2.o
From rdf_describedBy t1, rdf_qosprofilesubject t2
Where t1.s = 'WS1'
and t1.o = t2.i

=================================================================

Q6 Retrieve quality-of-service profiles of two Web services for comparison.

Q6 SPARQL

PREFIX geoseed: <http://geoseed.cs.panam.edu/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?profile ?pre ?obj
WHERE { {{geoseed:WS1 geoseed:describedBy ?profile .}
    UNION
    {geoseed:WS5 geoseed:describedBy ?profile .}}
    ?profile rdf:type geoseed:QoSProfile .
    ?profile ?pre ?obj . }

Q6 TRIPLE

Select t3.s, t3.p, t3.o
From rdf_Triple t1, rdf_Triple t2, rdf_Triple t3
Where (t1.s = 'WS1' or t1.s = 'WS5') and t1.p = 'describedBy' and t2.s = t1.o
and t2.p='rdf:type' and t2.o = 'QoSProfile' and t3.s = t2.s
Q6 PROPERTY

select x2.s, x2.o

from

(Select t2.s as s
From rdf_describedBy t1, rdf_type t2
Where (t1.s = 'WS1' or t1.s = 'WS5') and t2.s = t1.o
and t2.o = 'QoSProfile') x1

inner join {
select * from
rdf_type
union
select * from
rdf_trust
union
select * from
rdf_reliability
union
select * from
rdf_availability
union
select * from
rdf_processingTime
union
select * from
rdf_requestspersecond
union
select * from
rdf_security
union
select * from
rdf_failure
union
select * from
rdf_overallrating
union
}
Q6 PROFILE

Select t2.i, t2.p, t2.o
From rdf_describedBy t1, rdf_qosprofilesubject t2
Where (t1.s = 'WS1' or t1.s = 'WS5')
and t1.o = t2.i

Sample 1 Data set for Performance Study (GeneralProfile & QoSProfile)

<SP1> <rdf:type> <ScientificComputationalEntity> .
<SP1> <name> "gridding" .
<WS1> <rdf:type> <WebService> .
<GP1> <rdf:type> <GeneralProfile> .
<QoSP1> <rdf:type> <QoSProfile> .
<WS1> <describedBy> <GP1> .
<WS1> <describedBy> <QoSP1> .
<SP1> <implementedBy> <WS1> .
<GP1> <name> "name of a web service" .
<GP1> <author> "author1" .
<GP1> <contact> "contact1" .
<GP1> <contributor> "contributor1" .
<GP1> <description> "description1" .
<GP1> <url> "url1" .
<GP1> <identifier> "identifier1" .
<GP1> <version> "version1" .
<GP1> <releaseDate> "releaseDate1" .
<GP1> <language1> "language1" .
<GP1> <cost> "cost1" .
<GP1> <license> "license1" .
<GP1> <support> "support1" .
<QoSP1> <trust> "5" .
<QoSP1> <reliability> "1.0" .
<QoSP1> <availability> "0.8" .
<QoSP1> <processingTime> "250 ms".
<QoSP1> <requestsPerSecond> "40".
<QoSP1> <security> "0".
<QoSP1> <failure> "unknown".
<QoSP1> <overallRating> "5".
<QoSP1> <userReview> "User wrote something".

Sample 2 Dataset used in the Performance Study (GeneralProfile & QoSProfile with different descriptor set)

<SP2> <rdf:type> <ScientificComputationalEntity>.
<SP2> <name> "gridding".
<WS2> <rdf:type> <Application>.
<GP2> <rdf:type> <GeneralProfile>.
<QoSP2> <rdf:type> <QoSProfile>.
<WS2> <describedBy> <GP2>.
<WS2> <describedBy> <QoSP2>.
<SP2> <implementedBy> <WS2>.
<GP2> <name> "name of a application".
<GP2> <author> "author2".
<GP2> <contact> "contact2".
<GP2> <contributor> "contributor2".
<GP2> <description> "description2".
<GP2> <url> "url2".
<GP2> <identifier> "identifier2".
<GP2> <version> "version2".
<GP2> <releaseDate> "releaseDate2".
<GP2> <language2> "language2".
<GP2> <cost> "cost2".
<GP2> <license> "license2".
<GP2> <support> "support2".
<QoSP2> <reliability> "0.9".
<QoSP2> <availability> "1.0".
<QoSP2> <processingTime> "36 ms".
<QoSP2><requestsPerSecond>"100".
<QoSP2> <security> "0".
<QoSP2> <failure> "type 3 failure".
<QoSP2> <overallRating> "3".

118
Sample 3 Data set for Performance Study (GeneralProfile & QoSProfile with different descriptor list)

<SP3> <rdf:type> <ScientificComputationalEntity> .
<SP3> <name> "contourmap" .
<WS3> <rdf:type> <WebService> .
<GP3> <rdf:type> <GeneralProfile> .
<QoSP3> <rdf:type> <QoSProfile> .
<WS3> <describedBy> <GP3> .
<WS3> <describedBy> <QoSP3> .
<SP3> <implementedBy> <WS3> .
<GP3> <name> "name of a web service" .
<GP3> <author> "author3" .
<GP3> <description> "description3" .
<GP3> <url> "url3" .
<GP3> <identifier> "identifier3" .
<GP3> <releaseDate> "releaseDate3" .
<GP3> <language3> "language3" .
<GP3> <support> "support3" .
<QoSP3> <trust> "5" .
<QoSP3> <availability> "0.8" .
<QoSP3> <processingTime> "150 ms" .
<QoSP3> <requestsPerSecond> "300" .
<QoSP3> <overallRating> "4" .
Sample 4 Data set for Performance Study (GeneralProfile only)

<SP3> <rdf:type> <ScientificComputationalEntity> .
       <SP3> <name> "seismic analysis" .
       <WS3> <rdf:type> <WebService> .
       <GP3> <rdf:type> <GeneralProfile> .
       <GP3> <contributor> "contributor1" .
       <WS3> <describedBy> <GP3> .
       <SP3> <implementedBy> <WS3> .
       <GP3> <name> "name of a web service" .
       <GP3> <author> "author3" .
       <GP3> <description> "description3" .
       <GP3> <url> "url3" .
       <GP3> <identifier> "identifier3" .
       <GP3> <releaseDate> "releaseDate3" .
       <GP3> <language3> "language3" .
       <GP3> <support> "support3" .
CURRICULUM VITAE

Pearl Weaver Martin Brazier was born in Lancaster County, Pennsylvania, the ninth of 13 children. She was valedictorian of her class at Warwick High School in 1962 and received her Bachelor's Degree in Mathematics from Millersville University in 1966, the first in her family to attend college. She taught 4 years of High School Mathematics in Columbia, Pa. and one year of high school Mathematics in Milwaukee, Wisconsin. She completed a MS degree in Mathematics from The Ohio State University in 1971 and a MS degree in Computer Science and Applications from The Virginia Polytechnic and State University in 1981. She taught Mathematics part-time at the Ohio State University in 1972, before having two children. She taught Mathematics and Computer Science part-time at Virginia Tech from 1974 to 1981. In 1981, she took a lecturer position teaching Computer Science in the Department of Mathematics at Pan-American University, Edinburg TX, which became The University of Texas-Pan American in 1989.

She designed and implemented the Bachelor's Degree in Computer Science, founded the Computer Science Department, which she chaired from 1995-2001, and continues to serve as the Undergraduate Program Coordinator. She was largely responsible for getting the BSCS Computer Science degree accredited by CAC/ABET in 2001. During her tenure as chair, the department established a MS degree in Computer Science in 1996 and the MSIT, Master of Science in Information Technology in 2001. She was promoted to associate professor in 2001. In 2008, she became the Director of the Computer Engineering Program, which she helped to design and establish in 2007.
In 2002, she enrolled in the Ph.D. program in Computer Science at the University of Texas – El Paso, where she developed her research interest in Web 2.0 technologies and their application to the Geology domain. She completed the Ph.D. in 2010.

Her recent research has been with Dr. Ann Q. Gates and Dr. Artem Chebotko. Recent presentations and publications have been with Chebotko, Gates, Piazza, and Salayandia (2009) International Conference on Semantic Web and Web Services (SWWS 2009), and with Chebotko, Gates, and Salayandia Third International Workshop on Scientific Workflows (SWF 2009).


Permanent Address: 921 Tori Lane
Edinburg, TX 78539