

Vocabulary Integration Environment: VINE

Entorno de Integración de Vocabularios: VINE

Luis Bermúdez, PhD.
Monterey Bay Aquarium Research Institute, Monterey, CA, USA
bermudez@mbari.org

Recibido para revisión 26 de Marzo de 2007, aceptado 15 de Junio de 2007, versión final 31 de julio de 2007

Resumen—Interoperabilidad entre sistemas de información distribuidos requiere acuerdos de estándares de metadatos, protocolos, interfaces y vocabularios controlados. Tales acuerdos se basan en adopciones de estándares publicados por organizaciones como la Organización Internacional para la Estandarización (International Organization for Standardization - ISO) y el Open Geospatial Consortium (OGC). Estos estándares son generales y no regulan en totalidad los vocabularios controlados que se utilizan para anotar metadatos. Por consiguiente, cuando comunidades comparten metadatos, se encuentran con conflictos semánticos resultado de las heterogeneidades de los vocabularios controlados. Por ejemplo “elevación del agua”, “nivel del agua” y “altura del agua”, son distintos conceptos que semánticamente son equivalentes. Pare poder resolver estas heterogeneidades semánticas, se creo el programa VINE, Entorno de Integración de Vocabularios. Esta herramienta fue utilizada exitosamente en el taller Avanzando Vocabularios Controlados (Advancing Domain Vocabularies), organizado por el proyecto Interoperabilidad de Metadatos Marinos (MMI) en 2005. Esta herramienta se especializa en crear relaciones de tipo tesauros para mapear vocabularios controlados representados en grafos de Resource Description Framework (RDF). VINE también permite búsquedas de texto libre en los grafos. VINE es un plugin de Eclipse desarrollado en JAVA, siguiendo el paradigma de Modelo Vista Controlador (MVC). Se discutirá la arquitectura de la herramienta y el concepto de búsqueda inteligente profunda en grafos.

Palabras Clave—Ingeniería de Software, Aplicaciones de Inteligencia Artificial, Sistemas Orientados a Objetos.

Abstract—Interoperability between distributed information systems requires an agreement of metadata standards, protocols, interfaces and controlled vocabularies. Such an agreement is often pursued by adopting standards published by organizations such as the International Organization for Standardization (ISO) and the Open Geospatial Consortium (OGC). These standards are general and do not fully dictate the control vocabularies used for

annotating metadata. Therefore, when communities share metadata they encounter semantic conflicts because of the

heterogeneities of the control vocabularies. For example, “stage”, “gage height” and “water elevation” are different concepts that are semantically equivalent. To be able to solve these semantic heterogeneities, VINE, the Vocabulary Integration Environment tool, was created. This tool was successfully used at the workshop “Advancing Domain Vocabularies”, hosted by the Marine Metadata Initiative (MMI) in 2005. This tool specializes in “thesauri type” relations to map controlled vocabularies, encoded as (Resource Description Framework) RDF graphs. VINE also allows free text searches in the graph. It is a JAVA Eclipse Plugin and it follows the Model View Controller paradigm. The architecture of the tool and the concept of the smart deep graph search will be discussed.

Keywords—Computer Software, Artificial Intelligence, Software Tools.

I. INTRODUCTION

INTEROPERABILITY between distributed information systems requires an agreement of metadata standards, protocols, interfaces and controlled vocabularies. Such an agreement is often pursued by adopting standards published by organizations such as the International Organization for Standardization (ISO) and the Open Geospatial Consortium (OGC); however controlled vocabularies are domain specific, and semantic conflicts within information communities are found when extracting metadata. For example, *stage*, *gage height* and *water elevation* are different concepts that are semantically equivalent. As a consequence, information communities that share these concepts will have to resolve the semantic incompatibles to be able to exchange meaningful information. This is commonly known as the need to achieve semantic interoperability among heterogeneous systems [4, 9, 11, 12, 23]. The problem presents itself especially in *domain-specific metadata* values (e.g. land-cover, stage height, runoff), which are metadata that capture meaningful information specific to a domain [14].

Semantic heterogeneities among Information Communities could be solved by involving experts in the domain and allow them to perform relations between terms that are used by the information communities. The World Wide Web Consortium (W3C) has published a set of recommendations related to the Semantic Web [3] describing how to encode and relate concepts so that they can be understandable by distributed programs in the Web. One exemplar of an approach involving domain experts and the use of W3C technologies to solve semantic heterogeneities was the workshop “Advancing Domain Vocabularies” hosted by the Marine Metadata Initiative (MMI) in 2005. The workshop gathered domain experts to perform mappings among marine vocabularies. The tool used for the mappings relied on the Web Ontology Language (OWL) as the underlying model to manipulate the vocabularies and perform the relations among them. The tool is called VINE, and was developed in response to the lack of a mapping tool that would allow to search terms and map them in an easy fashion. About 10 sessions were hosted before the workshop to get requirements that drove the development and improvement of VINE. At the workshop VINE was very successful. It has become a popular tool to map controlled vocabularies to common standard ones.

VINE allows users to perform mapping between ontologies written for the Semantic Web. The Semantic Web is a W3C effort to allow computer programs to understand each other in the Web. The Semantic Web is build on standard formats such as XML and models such as the Resource Description Framework (RDF [5]). In the Semantic Web, ontologies, which are knowledge representations of a domain, are based on RDF.

This paper discusses why VINE is important to solve semantic heterogeneities between information communities, while incorporating domain experts and using simple thesauri-type mappings. Section 2 presents an introduction to the Semantic Web. Section 3 explains ontology integration and mapping. Section 4 discusses the *smart deep graph* search. Section 5 presents the architecture of VINE. Section 6 presents a comparison with the most popular ontology tools, and section 7 discusses performance and evaluation results.

II. SEMANTIC WEB

In computer science an ontology is an explicit and formal specification of mental abstractions, which conforms to a community agreement about a domain and design for a specific purpose [10]. It is different from the term Ontology (first letter in upper case) used in Philosophy to describe the existing things in the world [8]. Different abstractions, specifications and agreements exist among communities; so, different domain ontologies exist, while only a single Ontology is possible.

Ontology provides the structure of a controlled vocabulary similar to a dictionary or a thesaurus. The vocabulary agreed

to by a community is the expression of concepts (i.e., mental abstractions) of their domain. Since a concept can be expressed in different ways and differ in meaning from one person to another, the controlled vocabulary helps solve semantic incompatibilities [4, 11, 12, 23].

The Web Ontology Language, OWL [2] is a core component of the Semantic Web [3]. The Semantic Web is a universe of metadata and ontologies expressed in machine readable format, along with software tools that allow the understanding of semantic relations among heterogeneous and distributed resources in the Web [7]. OWL is based on technologies recommended by the World Wide Web Consortium, such as the extensible Markup Language (XML), Resource Description Framework (RDF), and Uniform Resource Identifier (URI). This last one allows a Web user to display a page by clicking on a link, download a file, or to name distinctly every resource in the Web. RDF and OWL use the URI to link, talk about, complement, use, and extend distributed resources.

RDF is based on statements that resemble simple language expressions. Statements are composed of a resource (*subject*) with a property (*predicate*) and a value (*object*). For example a statement could be “temperature is of type parameter”. In this case, *temperature* is the subject, *type* is the predicate and *parameter* is the object. Formally, the *subject* and *predicate* must be a URI. The *object* can also be any literal value and does not necessary need to be a URI. The previous example can formally be expressed as shown in

Table 1.

Subject and *objects* can be seen as nodes in a graph. An *object* can be a *subject* in another triple and a *subject* can be part of more than one statement. An ontology can be seen as a simple graph that can have thousands, or sometimes millions, of nodes.

Table 1. RDF Triple

Triple Part	Value
Subject	http://marinemetadata.org/2005/03/voc#temperature
Predicate	http://www.w3.org/1999/02/22-rdf-syntax-ns#type
Object	http://marinemetadata.org/2005/03/voc#parameter

III. ONTOLOGY INTEGRATION AND MAPPING

Ontology integration consists of bringing together two or more ontologies, expressing the result of such agreement in a new ontology [15, 18]. Integration is also known as merging [17]. In the integration process, two terms from two different ontologies are related by making an equivalent, subsumption, disjoint or instance relationship among them [17]. We define the term *mapping* similar to integration, as the process of creating a relation between two terms from different ontologies. The relations that are expressed in a new ontology can be equivalent, subsumption, disjoint, instance or any other.

To simplify the mapping of vocabularies, VINE uses three

default relations: *sameAs*, *narrowerThan* and *broaderThan*. The first relation is an OWL property, while the other two are VINE-defined. If needed, these properties can be set equivalent to SKOS [26] defined properties *exactMatch*, *narrowerMatch*, *broaderMatch*, respectively. Note that *sameAs* is a symmetric relation ($A = B$ implies $B = A$), while *narrowerThan* and *broaderThan* are inverse to each other ($A > B$ implies $B < A$) and are transitive ($A > B$ and $B > C$ implies $A > C$). This will allow inference of vocabulary relationships. An example of mapping the term bioluminescence among three controlled vocabularies using *owl:sameAs* is shown in Figure 1

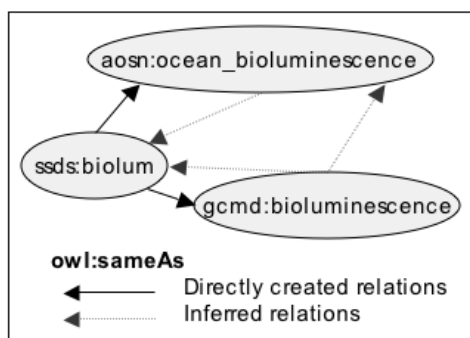


Figure 1. Same As relation and inference

IV. SMART DEEP GRAPH SEARCH

As said before, VINE opens ontology files that are based on RDF. In VINE, a resource (URI) is associated with a text, which is a combination of simplified statements. A simplified statement is a triple converted to a convenient string by removing the namespaces, removing the property and converting every letter to lower case. For example for the triple presented in Table 1, the simplified statement is "temperature parameter". This removes the complexity of parsing namespaces and searching on unhelpful properties.

Two types of searches are supported, a simple search and a REGEX (Perl 5-compatible regular expression) search. In the search field a user can write the terms to search separated by spaces. The user clicks on the search button or the return key. VINE will perform an AND search for all the terms given which are separated by spaces. A user could also enter a "-" before a word (without a space between the "-" and the word) to tell VINE to filter (exclude) that term. A history of the searched terms is kept so the user can select one of the history searches in the combo box.

Table 2. Query Examples

Query	Results
temp	tkel1 temperature in kelvin tke1ctd temperature in kelvin ctd tke2 temp k
ctd temp	tke1ctd temperature in kelvin ctd
temp -ctd	tkel1 temperature in Kelvin tke2 temp k

pera tke elvin	tkel1 temperature in kelvin tke1ctd temperature in kelvin ctd
----------------	--

To better understand how the search works, and example of queries and results are presented in Table 2. These examples assume that the following statements are the only statements in a given ontology:

- tkel1 temperature in kelvin,
- tke1ctd temperature in kelvin ctd
- tke2 temp k

V. VINE ARCHITECTURE

A. Open Source

VINE is developed as a Rich Client Platform (RCP) following the Eclipse plugin framework. An RCP application allows developers to add extra functionality within a well establish system architecture. Eclipse is based on the Standard Widget Toolkit (SWT) instead of the JAVA Abstract Windowing Toolkit (AWT). The SWT implementation uses the Java Native Interface (JNI) to access the native Graphical User Interfaces of each operating system. Even though SWT is coded in JAVA, RCP applications are not portable so an application must be tailored for a specific platform. However, the Eclipse IDE (Integrated Development Environment) provides a wizard that allows exporting the application to different application with very few clicks.

VINE source code and standalone applications for different platforms are available in SourceForge under the GPL license. This was a prerequisite for the development of the tool, since the MMI project is a community-centered initiative that as part of its mission provides useful and free tools to the community.

B. User Interface

VINE allows a user to select one or more ontologies for searching and mapping.

Figure 2 shows two popular controlled vocabularies in environmental systems in the international community. The CF, or Climate Forecast, COARDS convention parameters [6] and a the Global Change Master Directory Science Keywords, GCMD [20]. The former is selected in the left form and the later is selected in the right form. Within each form a user can search terms and find resources of interest. Detailed information about each resource is displayed in the bottom of the page.

The user can check resources from both left and right forms and perform a one-to-one, one-to-many, or many-to-one mapping using the property icons located in the middle of the page (e.g., '<'). New predicate terms can be created using the '?P' icon, and level of certainty for a mapping can be indicated with a percentage control. Information about mappings is giving in the middle of each form, and to view and modify all the created mappings the MappingResults

page can be selected.

C. VINE Main Composites

Each widget in VINE (Figure 3) is extended from a functional Eclipse SWT (Standard Widget Toolkit) widget. For example the *MultipageEditor* extends the SWT *MultiPageEditorPart*, which allows having access to multiple pages at a single time. The *MultipageEditor* contains two pages: a *MapperPage* and a *MappingResultsPage*. The *MapperPage* contains two *VocabularyForms*. Each *VocabularyForm* is a separate object where ontologies can be added and removed, and where searching is performed independently.

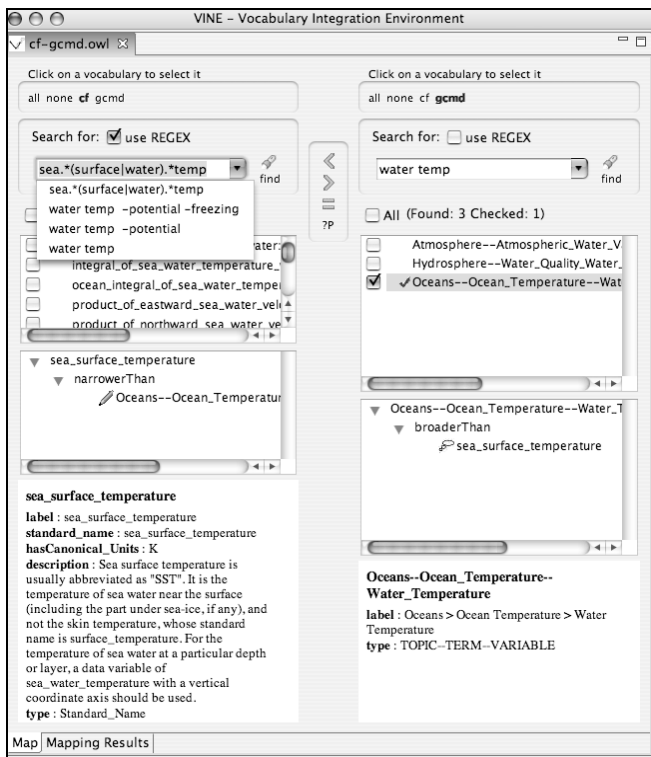


Figure 2. VINE graphical user Interface

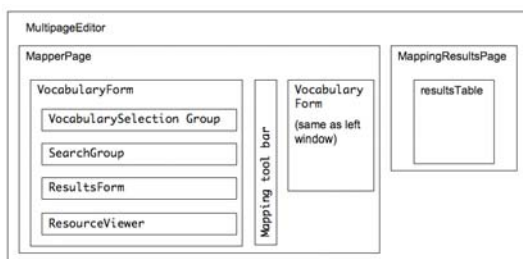


Figure 3. VINE main widgets

D. Main Interfaces

VINE is composed of an *ontologyModel*, a *MapperModel*, and a *PrefixManager*. The ontology model is a JENA [13] model, which is used to manage ontology objects. The

mapperModel is a subclass of the *ontologyModel*. The *mapperModel* relies on a *MappingPolicy* instance to determine if a relation (statement) is classified as a mapping or not. The *MappingPolicy* could be implemented in various ways depending on the mapping strategy (e.g. MMI strategy, SKOS [26] or any other). The *mapperModel* contains listeners that are notified any time the *MapperModel* state changes. Figure 4 shows these components schematically.

The *ModelSearcher* is an interface that searches a *MapperModel*. Classes representing different searching strategies implement the *ModelSearcher* interface, such as *MMISmartDeepGraphSearcher* and *PennStateCorpusStateMatcher*. This last one is a VINE plugin that allows users to search related terms by using an underlying ontology such as Wordnet. The concepts in the ontology are linked as nodes to an RDF representation of Wordnet. When a search is performed about a term, a concept in Wordnet is found and the closest nodes corresponding to concepts in the ontology are returned.

E. Model View Controller

VINE is designed to follow the Model View Controller (MVC) pattern (

Figure 5). As discussed before, it uses SWT composites to create its views. The views are *MapperModelListeners* that get notified every time the state of the *MapperModel* changes. All the widgets in the mapping page (Figure 3) are *MapperModelListeners*.

The *MapperPageController* gets gestures from users, performs changes in the *MapperModel*, and triggers a specific view depending on the user action. Views do not directly change the model. If a view wants to receive notification from the *MapperModel* it needs to register to the *MapperModel*, which is responsible for adding, removing and notifying the listeners.

VI. COMPARISON WITH OTHER TOOLS

VINE is a specialized tool and differs with other ontology editors, in that its searching capability is more sophisticated and it has three simple mapping defaults (*sameAs*, *broaderThan* and *narrowerThan*) that can be applied to imported ontologies. If only subsumption relations (class-subclass) are required to achieve integration between ontologies, other tools like PROTÉGÉ [24] and SWOOP [16] will do this job better. PROTÉGÉ is a popular tool developed by the Stanford Medical Informatics group, while SWOOP is developed by MINDSWAP.

When performing free text searches, VINE searches for every resource in its graphs using regular expressions. Each resource is associated with a text that corresponds to the concatenated text of all the triples that include this resource. This search capability of VINE is not available in any other ontology editors, like PROTÉGÉ, SWOOP, POWL [1], OiLED [25], IsaViz [21]. The aforementioned tools provide limited support for searching only in the local name of the resources. However, some tools, like PROTÉGÉ, are starting

to support RDF query languages like SPARQL[22], which search queries could be complex for domain experts.

Thus, if the local name of a resource is a not human readable text, and the information (e.g. *rdf:comment*) about that resource is a literal value, SWOOP and PROTÉGÉ will not find this information. In addition, VINE will return matches from other triples that relate to this search term, extending the information made available to the user.

PROTEGE and SWOOP will both open one or more ontologies at the same time. To open more than one ontology, an ontology must import one or more ontologies. This is done by specifying in the ontology tag an import

element containing the URIs of each imported ontology. However, PROTÉGÉ and SWOOP only allow subsumption relations, between two concepts from imported ontologies, VINE, allows creating any type of relation (e.g. *narrowerThan*) between resources of imported ontologies.

There are other tools available for mapping. One is PROMPT [19] from Stanford University. This is a PROTÉGÉ [24] plugin and it allows users to compare and merge (suggested) mappings between two ontologies. However, it doesn't provide the searching capabilities of VINE and the ability to create relations described above.

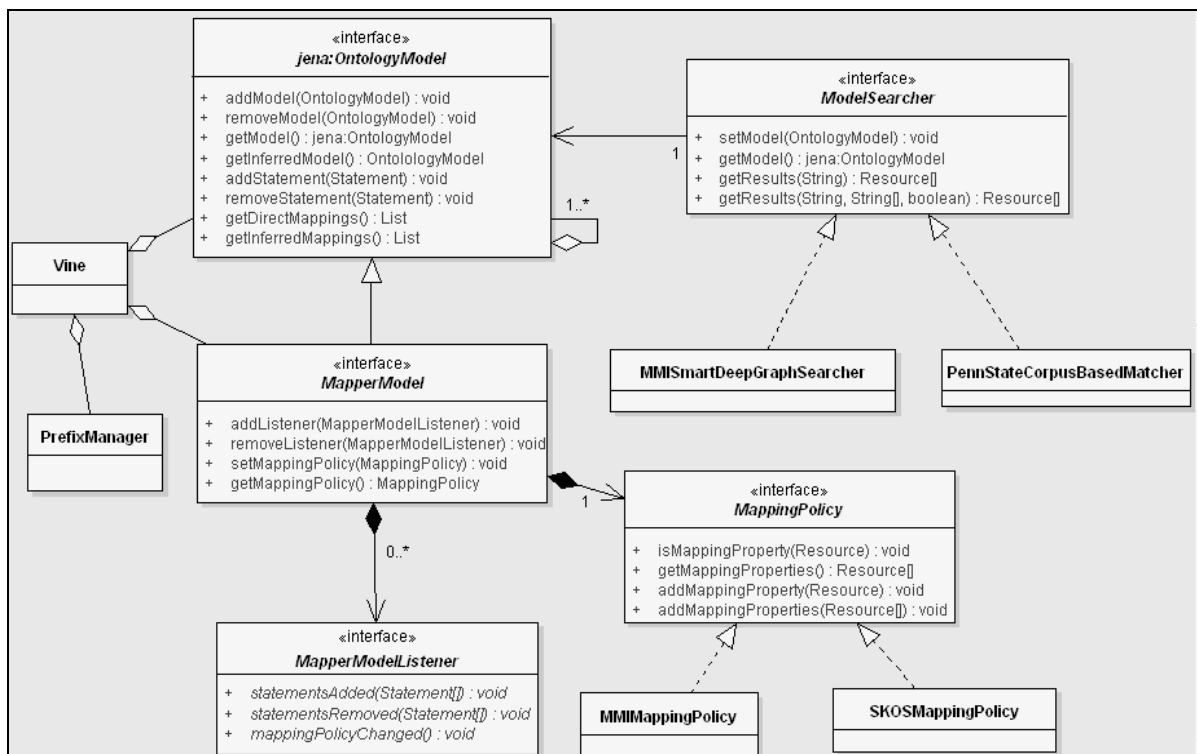


Figure 4. VINE main interfaces

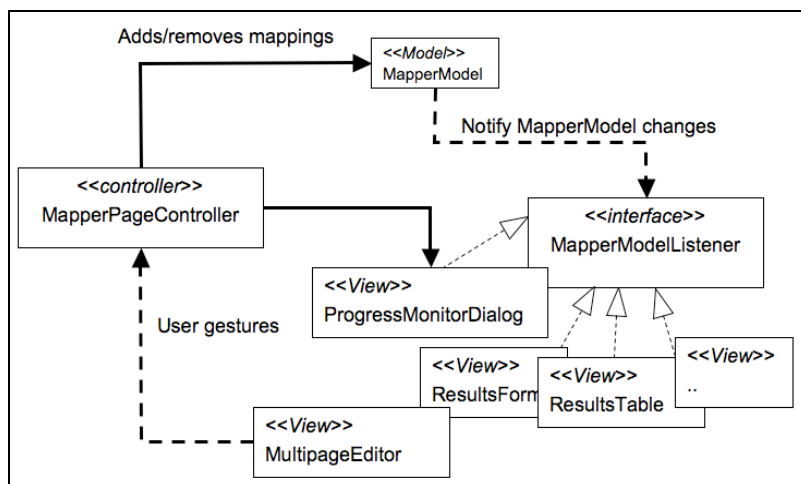


Figure 5. Model view controller in VINE

VII. PERFORMANCE AND WORKSHOP RESULTS

VINE proved a robust tool in the MMI workshop “Advancing Domain Vocabularies”, where it was used to map terminologies in a subset of marine domains. A summary of the number of mappings done by each 6-person group in 8 hours of mapping time is presented in Table 3. The participants had little problem running VINE with Windows, Macintosh and Linux operating systems. Also VINE proved able to handle hundreds of thousand of triples; for example, some groups opened up to 15 ontologies at a time. Details about these ontologies are available at <http://marinemetadata.org/allont>.

VINE has also been used to search for terms in a large database of marine data, and relate them to standard terms in a vocabulary—in this case, the COARDS Climate/Forecast (CF) vocabulary[4]. After converting the database into OWL RDF form, VINE enabled the synthesis of several hundred relationships from tens of thousands of terms. Originally projected to take multiple weeks, the entire mapping task took less than 8 hours.

Table 3. Mapping results

Topic	Direct mapping	Inferred mapping	Total mapping
Plant Pigments	405	1,022	1,427
PaCOOS	131	375	506
Waver	93	181	274
Currents	90	153	243
CTD	81	432	513
Habitats	23	37	60
Total	823	2,200	3,023

A future version of VINE will allow direct database connections and (with the help of a wizard) and automatic conversion from an SQL result set to OWL.

VIII. CONCLUSIONS

Solving semantic heterogeneities is necessary to achieve semantic heterogeneity among information communities. VINE is a tool that allows users to search efficiently and to very quickly make simple thesauri-type mappings. It was used successfully in a vocabulary mapping workshop, as well as for other tasks. And, even though it was used in the marine domain, the technology can be used in other domain where interoperability needs to be achieved. The architecture of the tool was presented, showing an Eclipse Plugin creation following the Model View Controller architecture and allowing plug-ins of other searching schemes.

VINE leverages Semantic Web technologies such as RDF and JENA API [13] to perform mappings between domain

vocabularies. Because of its sophisticated search capabilities, highly optimized user interface, and ability to relate term instances, VINE offers key advantages over existing tools.

ACKNOWLEDGMENT

This activity was supported by the National Science Foundation (grant ATM-0447031). The author gratefully acknowledges the support of the Monterey Bay Aquarium Research Institute (MBARI) and the David and Lucile Packard Foundation.

REFERENCES

- [1] Agile Knowledge Engineering and Semantic Web (AKSW), "POWL-Semantic Web Development Plattform," 2007.
- [2] S. Bechhofer, F. v. Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. Patel-Schneider, and L. A. Stein, "OWL Web Ontology Language Reference," W3C Recommendation 2004.
- [3] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, vol. 184, pp. 34-43, 2001.
- [4] Y. Bishr, "Overcoming the semantic and other barriers to GIS interoperability," *Geographic Information Science*, vol. 12, pp. 299-314, 1998.
- [5] D. Brickley and R. V. Guha, "RDF Vocabulary Description Language 1.0: RDF Schema," W3C 2004.
- [6] CF Standard Names, "2006."
- [7] M. J. Egenhofer, "Toward the Semantic Geospatial Web," presented at Tenth ACM international symposium on advances in geographic information systems, McLean, Virginia, USA, 2002.
- [8] F. T. Fonseca, "Ontology-Driven Geographic Information Systems," in *Spatial Information Science and Engineering*. Maine: The University of Maine, 2001.
- [9] C. H. Goh, "Representing and Reasoning about Semantic Conflicts in Heterogeneous Information Sources." Cambridge, Massachusetts: MIT Sloan School of Management, 1997.
- [10] T. Gruber, "A Translation Approach to Portable Ontology Specification," *Knowledge Acquisition*, vol. 5, pp. 199-220, 1993.
- [11] T. Hadzilakos, G. Halaris, M. Kavouras, M. Kokla, G. Panopoulos, I. Paraschakis, T. Sellis, L. Tsoulos, and M. Zervakis, "Interoperability and Definition of a National Standard for Geospatial Data: The Case of the Hellenic Cadastre," *International Journal of Applied Earth Observations and Geoinformation*, vol. 2, pp. 120-128, 2000.
- [12] F. Harvey, W. Kuhn, H. Pundt, and Y. Bishr, "Semantic interoperability: A central issue for sharing geographic information," *The Annals of Regional Science*, vol. 33, pp. 213-232, 1999.
- [13] HP Labs Semantic Web Research, "JENA," 2006.
- [14] V. Kashyap and A. Sheth, *Information Brokering Across Heterogeneous Digital Data*. Norwell, MA: Kluwer Academic Publishers, 2000.
- [15] M. Klein, "Combining and Relating Ontologies: An Analysis of Problems and Solutions," presented at Workshop on Ontologies and Information Sharing, IJCAI'01, Seattle, USA, 2001.
- [16] Maryland Information and Network Dynamics Lab Semantic Web Agents Project (MINDSWAP), "SWOOP," 2006.
- [17] D. L. McGuinness, R. Fikes, J. Rice, and S. Wilder, "An Environment for Merging and Testing Large Ontologies," presented at Principles of Knowledge Representation and Reasoning (KR2000), San Francisco, USA, 2000.
- [18] E. Mena, V. Kashyap, A. Illarramendi, and A. P. Sheth, "Managing multiple information sources through ontologies: relationship between vocabulary heterogeneity and loss of information," presented at 3rd Workshop Knowledge Representation Meets Databases (KRDB'96), Budapest, Hungary., 1996.
- [19] N. Noy, "PROMPT," 2005.

- [20] L. M. Olsen, G. Major, S. Leicester, K. Shein, J. Scialdone, H. Weir, S. Ritz, C. Solomon, M. Holland, R. Bilodeau, T. Northcutt, and T. Vogel, "Global Change Master Directory (GCMD) Earth Science Keywords," NASA, 2006.
- [21] E. Pietriga, "IsaViz," 2007.
- [22] E. Prud'hommeaux and A. Seaborn, "SPARQL Query Language for RDF," W3C 2006.
- [23] A. P. Sheth, "Changing focus on interoperability in information systems: from system, syntax, structures to semantics.," in *Interoperating geographic information systems*, M. F. Goodchild, M. J. Egenhofer, R. Fegeas, and C. Cottman, Eds. Boston: Kluwer Academic Publishers, 1999, pp. 5-29.
- [24] Stanford Medical Informatics, "Protege," 3.2 Beta ed, 2006.
- [25] University of Manchester, "OilED," 2007.
- [26] World Wide Web Consortium, "Simple Knowledge Organization System (SKOS)," 2005.

Luis E. Bermudez (M'06) born in Bogota, Colombia, 1971, received his B.S. from Universidad de los Andes, Bogota, Colombia in Industrial Engineering and an MS. and PhD. from Drexel University, Philadelphia, PA, in Hydroinformatics.

He is a Software Engineer at the Research and Development department of the Monterey Bay Aquarium Research Institute, at Monterey CA. He is technical leader of the Marine Metadata Interoperability Initiative, and serves in various committees related to integration of observing systems. He has authored and coauthor more than 15 papers related to knowledge representation and information modeling for geosciences, metadata and data for geographic information systems and the sensor web. Most recent publication: Bermudez, L.E., and Piasecki, M., "Metadata Community Profiles for the Semantic Web" *Geoinformatica*, 2006, 10(2), pp. 159-76.

Dr. Bermudez is member of Geological Society of America, where he serves in the Geoinformatics group, the American Geophysical Union, and the Association for Computer Machinery.

Universidad Nacional de Colombia Sede Medellín Facultad de Minas



Escuela de Ingeniería de Sistemas

Pregrado

- ❖ Ingeniería de Sistemas e Informática.



Áreas de Investigación

- ❖ Ingeniería de Software.
- ❖ Investigación de Operaciones.
- ❖ Inteligencia Artificial.

Posgrado

- ❖ Doctorado en Ingeniería-Sistemas.
- ❖ Maestría en Ingeniería de Sistemas.
- ❖ Especialización en Sistemas con énfasis en:
 - Ingeniería de Software.
 - Investigación de Operaciones.
 - Inteligencia Artificial.
- ❖ Especialización en Mercados de Energía.

Escuela de Ingeniería de Sistemas
Dirección Postal:
Carrera 80 No. 65 - 223 Bloque M8A
Facultad de Minas. Medellín - Colombia
Tel: (574) 4255350 Fax: (574) 4255365
Email: esistema@unalmed.edu.co
<http://pisis.unalmed.edu.co/>

