Departmental Technical Reports (CS)   Computer Science

7-2007

The Gravity Data Ontology: Laying the Foundation for Workflow-Driven Ontologies

Ann Q. Gates
The University of Texas at El Paso, agates@utep.edu

G. Randy Keller
The University of Texas at El Paso, keller@utep.edu

Flor Salcedo

Paulo Pinheiro da Silva
The University of Texas at El Paso, paulo@utep.edu

Leonardo Salayandia
The University of Texas at El Paso, leonardo@utep.edu

Comments:
Technical Report: UTEP-CS-07-44

Recommended Citation
Gates, Ann Q.; Keller, G. Randy; Salcedo, Flor; Pinheiro da Silva, Paulo; and Salayandia, Leonardo, "The Gravity Data Ontology: Laying the Foundation for Workflow-Driven Ontologies" (2007). *Departmental Technical Reports (CS)*. 185.
https://scholarworks.utep.edu/cs_techrep/185

The Gravity Data Ontology: Laying the Foundation for Workflow-Driven Ontologies

Ann Q. Gates, G. Randy Keller, Flor Salcedo, Paulo Pinheiro da Silva, Leonardo Salayandia

The University of Texas at El Paso; University of Oklahoma; Rockwell Collins
agates, paulo, leonardo@utep.edu; grkeller@ou.edu; fisalced@rockwellcollins.com

Abstract. A workflow-driven ontology is an ontology that encodes discipline-specific knowledge in the form of concepts and relationships and that facilitates the composition of services to create products and derive data. Early work on the development of such an ontology resulted in the construction of a gravity data ontology and the categorization of concepts: "Data," "Method," and "Product." "Data" is further categorized as "Raw Data" and "Derived Data," e.g., reduced data. The relationships that are defined capture inputs to and outputs from methods, e.g., derived data and products are output from methods, as well as other associations that are related to workflow computation. This paper describes the construction of a workflow-driven ontology that documents the methods and processes associated with gravity data and related products. In addition, the paper describes the progress done on the process to create workflow-driven ontologies, such that scientists are supported to create and validate such ontologies, while still enabling the automatic generation of executable workflow applications.

Keywords: ontology, workflow, workflow-driven ontology, gravity data, geospatial data

Introduction

Numerous institutions and organizations around the country have collected geospatial data and algorithms and processes for manipulating and integrating these data with other diverse data sets, generating results that are useable by them, other scientists, or the general public. The goal of the work presented in this paper is to move from an environment in which a scientist relies on a professional network and manual processes to complete their work to one in which a scientist uses an automated system to complete tasks or obtain results using knowledge from one or more domain experts. A particular area of concern is capturing knowledge in a particular discipline through an ontology and leveraging the knowledge to support the design and execution of

scientific workflows that compose software services to compute a particular result or generate a product.

There are several challenges that scientists face when creating any ontology: defining the scope of knowledge capture, determining the level of abstraction used to describe concepts and relationships, and identifying useful concepts and relationships. Clearly, creation of an ontology should be a continuing process that requires revision and refinement.

This paper introduces the notion of a workflow-driven ontology, one in which discipline-specific knowledge is encoded in the form of concepts and relationships that support visualizations depicting how data is derived or results are obtained, e.g., in the form of a workflow. In addition, the paper describes the construction of such an ontology for gravity data that documents the methods and processes associated with gravity data and related products. Last, the paper describes the progress done on the process to create workflow-driven ontologies [10], [11] and presents an overview of an effort to develop tools that assist a scientist during the process of creating and validating an ontology and generating abstract workflows. These workflows denote how a result is achieved by presenting the composition of methods (software services or algorithms) including the flow of data and control among the methods.

Motivation

The basis for the concept of a workflow-driven ontology was inspired by a February 2004 Seismology Ontology workshop held at Scripps Institution in San Diego. The attendees of the workshop included experts in the areas of seismology and information technology[1].

While the initial focus of the workshop was on creating a discipline-based ontology, i.e., an ontology focused on capturing knowledge about a particular discipline, it ended with a categorization and a set of relationships that were based on a general workflow that describes a common task performed by seismologists. After struggling with identifying the concepts that should be captured in a seismology ontology and motivated by a desire to identify concepts and relationships that would be useful to the community, the workshop participants defined concepts of interest by constructing the

---

[1] Randy Keller and Ann Gates, University of Texas at El Paso; Bertram Ludaescher, Dogan Seber, Chaitan Baru, and Kai Lin, San Diego Supercomputer Center; Gabi Laske and Frank Vernon, Scripps Institute, University of California at San Diego; Tim Ahern, IRIS; Colin Zelt, Rice; Matt Fouch, Arizona State; John Hole, Virginia Tech; David James, Carnegie Institute of Washington; and Bill Pike, Penn State.

workflow shown in Fig. 1. For the scientists, the workflow captured the steps for completing the task of creating a P-wave velocity model, and the necessary concepts that are involved in completing such a task. After completing the workflow, the seismologists next partitioned the diagram to into three categories: "Data," "Method," and "Product," where *Data* denotes input to or output from a *Method*, *Method* is a software service or algorithm, and a *Product* is an artifact.
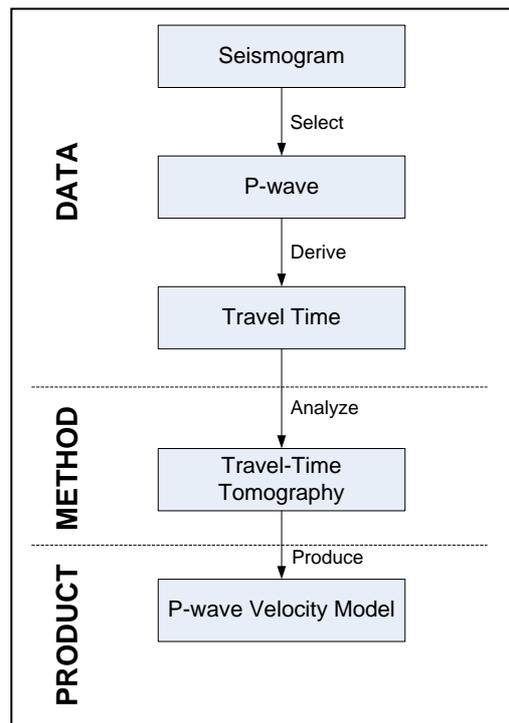

Fig. 1. A workflow created at the 2004 Seismology Ontology Workshop and used to define the concepts or classes of an ontology.

A summary of observations from the workshop includes the following:
1. *The benefits of using a workflow to drive creation of an ontology-* If one considers how a desired product or result is generated, a discipline expert can identify the data, derivation algorithms, transformation algorithms, and other data processing algorithms involved as well as the relationships between them. The capture of these concepts in an ontology can help scientists document the essential knowledge and processes required to advance their field.
2. *The benefits of using a workflow to determine missing concepts or relationships-* It's important to note that the workflow given in Fig. 1 is not

complete. The step from *P-Wave* to *Travel Time* requires a transformation method that is not depicted in the diagram. The ability to view a workflow based on concepts captured in an ontology can assist in the iterative process of refining an ontology.

3. *The importance of using abstraction in the ontology-construction process.* Related to the second observation, observation three promotes the need to focus on a particular product or result at a high-level while neglecting other aspects. On subsequent iterations devoted to refining the ontology, other aspects become the focus. Moving from a high-level abstraction to detail allows one to manage the complexity in defining an ontology. For example, one can specify that *P-Wave derives Travel Time* and in subsequent iterations specify the method by which this is done. As another example, consider that one specifies that one method includes several others. Further refinement would focus on the temporal aspects of this relationship.

4. *The importance of having ontologies that are created by scientists and for scientists.* While technology is critical for the development of cyberinfrastructure, the tools that scientists use to define and manage ontologies and workflows must be scientist-friendly and relevant to them.

Workflow-Driven Ontologies

The Notion of a Workflow-Driven Ontology

The observations that were made at the 2004 Seismology Workshop led to the definition of a specialized ontology called a *workflow-driven ontology,* (originally named as a *computation-driven ontology*), an ontology that encodes discipline-specific knowledge in the form of concepts and relationships supporting visualizations that depict how data is derived or results are obtained, e.g., in the form of a workflow. A workflow-driven ontology categorizes concepts and uses a set of workflow-related relationships.

As a proof-of-concept, Salcedo and Keller [12] applied the approach to develop a gravity-data ontology. The top-level categories of the ontology are described as they apply to the gravity domain:

- *Data* includes value type information related directly to a specific field of study. For example, for the gravity data domain there are three types of data: (1) Field Observations, the purest form of gravity data; (2) Principal Facts, i.e., latitude, longitude, elevation and observed gravity values; and (3) Derived (Reduced) Data, i.e., values that are perceived and sought as data by the user community. All three types are values associated with a point.

- *Methods* are algorithms that are applied to the various forms of data to produce results that are interpretable from a geologic point of view. Results from methods yield derived data or products.
- *Products* are artifacts that result from application of a method. These artifacts are not perceived and sought as data by the user community. Examples include maps, models, or images.

Table 1 summarizes the main relationships that are defined for a workflow-driven ontology. The table gives the inverse relationships and indicates whether the relationship supports transitivity, i.e., if *a* is related to *b* and *b* is related to *c*, then *a* is related to *c*.

Table 1. A summary of relationships for a geospatial computation-driven ontology.

TUPLE	INVERSE	TRANS.	DESCRIPTION
<c1, isInputTo, c2>	getsInputFrom	No	c1 is a Data or Product with raw numerical values concept; c1 is input into Method c2
<c1, isOutputOf, c2>	outputs	No	c1 is a Data or Product concept c2 is a Method concept
<c1,isDerivedFrom, c2>	isConvertedTo	Yes	c1 is a Data or Product concept c2 is a Data or Product concept c1 has been created through a transformation of c2 c1's existence depends upon the existence of c2.
<c1, includes, c2>	isIncludedIn	Yes	A Method c1 includes a Method c2 as a helper method.
<c1, uses, c2>	isUsedFor	Yes	c1 is a Method concept c2 is a Product or Data A method uses a product or data when neither one is direct input into the method.

Consider the following statement: the adjusted gravity reading in milligals is derived from the raw gravity reading via the equation:

$$AGR= (RGR * CC) + DC + TC,$$

where *AGR* is the adjusted gravity reading, *RGR* is raw gravity reading, *CC* is calibration constant for the gravity meter, *DC* is drift correction, and *TC* is tidal correction. From this text, we identify a method M_{AGR} that computes *AGR,* and we identify the following relationships:

- $<RGR$, isInputTo, $M_{AGR}>$
- $<CC$, isInputTo, $M_{AGR}>$
- $<DC$, isInputTo, $M_{AGR}>$
- $<TC$, isInputTo, $M_{AGR}>$
- $<AGR$, isOutputOf, $M_{AGR}>$

In the initial iteration of the ontology, one could state: *<AGR*, isDerivedFrom, *RGR>*, if the equation was not available or not considered because that level of detail was being abstracted.

The next example shows the application of the *include* relationship, and makes an argument for incorporating it in a workflow-driven ontology. Consider the text: Gridding methods include interpolation methods. This could be denoted as: $<M_{Grid}$, *includes,* $M_{Inter}>$. There are a number of interpolation algorithms that could be used with a gridding algorithm, and the *includes* relationship is used to capture this notion.

To illustrate the *uses* relationship, consider the following statement: a Regional Gravity Map (*RGM*) is used to determine whether to use a Directional Filter Method because the user must visualize the anomaly values to decide whether to use this filter. This denotes a manual process and should be considered when deriving a workflow description. The relationship would be expressed as: $<M_{Filter,}$ *uses, RGM>*.

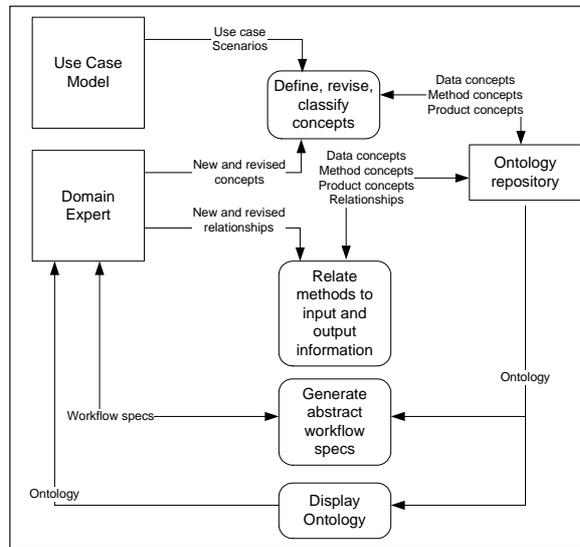**The Process of Constructing a Workflow-Driven Ontology**

Fig. 2. Flow of information when constructing a computation-based ontology.

Ontology 101: A Guide to Creating Your First Ontology [8] presents guidelines for creating an ontology, which are applicable to a workflow-driven ontology. In particular, use case modeling is an effective approach for driving the creation of any ontology.

The workflow-driven ontology approach places the primary focus on methods and data that generate results of interest to the scientist as well as on workflow-based relationships.   Fig. 2 presents a dataflow diagram that depicts the processes or steps for defining a workflow-driven ontology. The square in the diagram represents a source or sink, the rounded boxes depict transformation of information, and the open rectangle a store. As depicted in the figure, creation of a workflow-driven ontology (or any ontology) is a continuing process, and it includes the use of an abstract workflow (as depicted in Fig. 3). The processes are described next.

Identify concepts. Use cases allow one to scope the knowledge capture and identify useful concepts. In use-case modeling, the scientist identifies the primary uses of the ontology. Identifying use cases is complementary to developing workflows as an initial approach for specifying appropriate concepts. The discipline expert should consider the following questions: What types of data are available or can be derived? What existing algorithms, tools, or steps are used to generate data? What results are important to me or the community?

To illustrate the benefit of including use-cases in construction of a workflow-driven ontology for the gravity data domain, consider the following use cases: "determine the Complete Bouguer Anomaly for points in a gravity data set," and "create a free-air anomaly map." Given the use case as a starting point, the scientist would identify related algorithms for generating the desired data or product. For example, starting with the concept *Complete Bouguer Anomaly* and knowing that "Variations in Simple or Complete Bouguer Anomaly values are the major input into Interpretations of the geological features present in the area of a geophysical study" would lead to the following concepts (types in parenthesis): *Simple Bouguer Anomaly* (*Derived Data*), *Complete Bouguer Anomaly* (*Derived Data*), *Interpretation Method* (*Method*). The following statement, "Calculation of the Complete Bouguer Anomaly uses the Free Air Correction value," leads to the following concepts: *Calculate Complete Bouguer Anomaly* (*Method*) and *Free Air Correction Value* (*Derived Data*). The following statement, "Observed Gravity Data is input to the Calculate Free Air Anomaly method where it has modifications performed on it and this produces a Free Air Anomaly," leads to the following concepts: *Observed Gravity Data* (*Processed Data*), *Calculate Free Air Anomaly method* (*Method*), and *Free Air Anomaly* (*Processed Data*).

To elucidate the process of using a workflow to drive elicitation of concepts, consider that a discipline expert identifies *Anomaly Map* as an important result. Geospatial-mapping software, such as GMT (Generic Mapping Tools) [16] and denoted in the figure as *Mapping,* takes *Anomaly values,* grids them, and contours them to generate an *Anomaly Map. Anomaly values* are the result of raw gravity data reduction (e.g., [4]), which can be

obtained through a series of steps programmed in Excel (e.g., [5]). In this example, *Anomaly Map* would be classified as *Product* and *Anomaly value*s would be classified as *Derived Data*. *Mapping* and *Excel Reduction* are classified as *Methods*. Fig. 3 presents two views for specifying this workflow. In the first depiction, methods are shown on the right side of the diagram, data and products are shown on the left. The relationships are marked above the arrows. In the second, the text in bold denotes the desired output. Questions regarding "how the output is generated" results in the specification of the next step. This continues until the base or initial concept is reached, i.e., *Raw gravity data*. The darkened arrows denote the outputs from methods and the text within parenthesis denote the inputs to the methods.
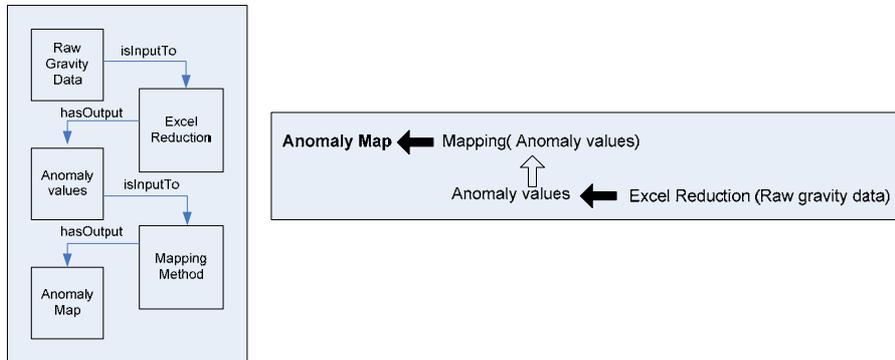

Fig. 3. Two views for illustrating the steps toward generating an abstract workflow specification for an Anomaly map.

Defining a simple workflow as shown in Fig. 3 can be useful for defining concepts as well as refining concepts. For example, if the discipline expert had not included the *Excel Reduction* method and instead used the relationship *Anomaly Values isDerivedFrom RawGravityData* in the first diagram of Fig. 3, then the expert would recognize that the ontology is underspecified; he or she would specify the method *Excel Reduction* during refinement.

    Identify relationships. The discipline expert also identifies the relationships between concepts. All *Derived data* and *Product* concepts should be associated with at least one *Method* class, and all *Method* classes should have input and output relationships.

    When appropriate, the domain expert defines a hierarchy of concepts, i.e., a structure of concepts in which C_i is of type C for each $I \in \{1,…,n\}$ as shown in Fig. 4. The gravity data ontology is represented in the Ontology Web Language (OWL) [14], and the concepts described in this paper are

referenced as classes in OWL. As a result, the class hierarchies are grounded in the OWL class *Thing*. During construction of the gravity data ontology, superclass *Product* was divided into subclasses *Gravity map* and *Gravity model*, and subclasses *Anomaly Map* and *Contour Map* were defined under *Gravity Map*.
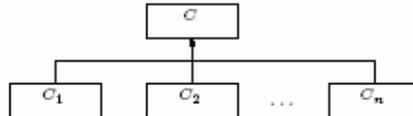

Fig. 4. A class hierarchy.

As described earlier, creation of an ontology should be a continuing process that requires revision and refinement. For example, refinement of the ontology resulted in refining the *Interpretation* concept to include subclasses *Modeling* and *Mapmaking*. A similar refinement process occurred in which concepts *Complete Bouguer Anomaly* and *Free Air Anomaly* were classified as *Corrected Gravity Data*.
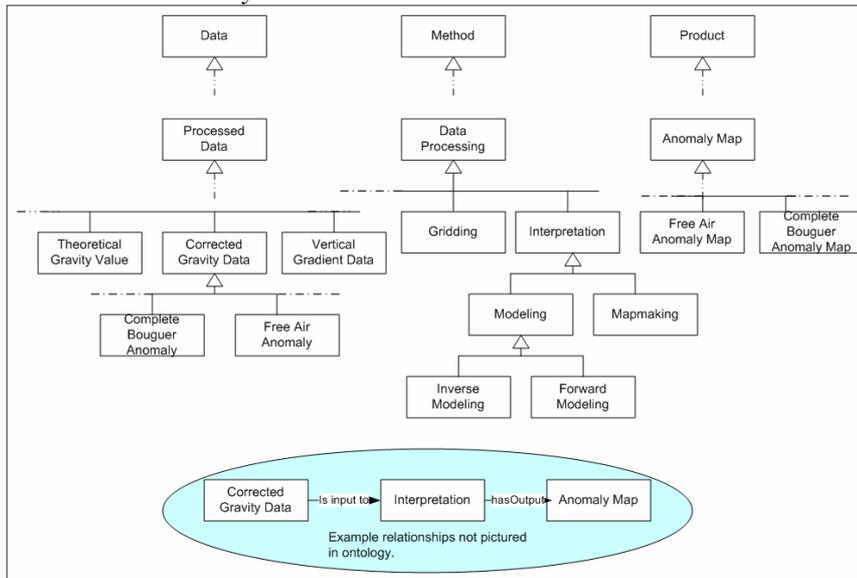
Fig. 5. Graphical representation of a portion of the Gravity Data WDO.

Gravity Data Ontology

Fig. 5 shows a portion of the gravity data ontology that was created with experts in the field of geophysics using Protégé Ontology Editor and Knowledge-Base Framework tool, Version 3.1 Beta Full. Because of space constraints, the graphical depiction does not show relationships or annotations associated with each concept. See http://trust.utep.edu/ciminer/collaborations/ for documentation of the ontology.

The Next-Generation Workflow-Driven Ontology

The experience of creating a workflow-driven ontology for gravity data provided a number of insights. The scientist involved in defining the gravity data ontology found it more amenable to work on an Excel worksheet to initially store the concepts and relationships prior to specifying them in Protégé. Moving toward a scientist-friendly approach to specification of ontologies has become a focus of the research.

As described in the previous section, workflows are essential in the workflow-driven approach to constructing ontologies. As such, it is important to be able to automate derivation of abstract workflows from an ontology to drive refinement and validation of the ontology. This will be an important step in assisting the refinement of the gravity-data ontology. A set of tools are under development to support the construction of the next generation of workflow-driven ontologies. The base hierarchy and relationships of the initial workflow-driven ontology are being revised to support the automated generation of abstract workflows, allowing the scientist to use the WDO to extract abstract workflow specifications.

The abstract workflow specifications generated from the WDO are useful in the refinement phases of ontology development as well as in creating executable workflow applications by supporting the Software Engineering principle of Separation of Concerns. Abstract workflow specifications allow scientists to craft requirement specifications without loosing sight of the task at hand, addressing the concern of the scientist. On the other hand, the abstract workflow specified by the scientist serves as a guide for the technologist, allowing her/him to focus on the appropriate details of the workflow specification to create an executable implementation, e.g., temporal

dependencies between workflow steps, addressing the concern of the technologist.

Abstract workflow specifications are generated with the aid of a tool called WDO-It! that interprets the knowledge represented in a WDO. The WDO-It! tool simplifies the capture and refinement of the ontology by presenting the concepts using a navigation tree structure that displays them as classes and subclasses, supporting the specification of input and output relationships for methods and data, and presenting a visualization of the abstract workflow specifications. The authors are in the process of validating the usability of the tool.

The scientist identifies the output information desired from the WDO, and WDO-It! then builds an abstract workflow specification based on the concepts and relationships defined in the WDO. In addition, the capture of provenance information [8] provides the scientist with the ability to annotate *Raw Data* with source metadata. For example, metadata regarding raw gravity data could include information about the instrument used to collect the raw data, accuracy estimates, and the individual or entity that recorded the readings; and the scientist can annotate *Processed Data* with method metadata, e.g., metadata about the method generating the processed data.

Related Work

There are numerous published ontologies. This section summarizes three: the Gene Ontology (GO), the Transparent Access to Multiple Biological Information Sources (TAMBIS), and the Semantic Web for Earth and Environmental Terminology (SWEET) ontologies. Two of the chosen ontologies have similarities with the workflow-driven ontology approach described in this paper.

The Gene Ontology (GO) [15] provides a controlled vocabulary to capture gene information. It is split up into three main categories: the cellular component ontology, the molecular function ontology, and the biological process ontology. In the GO ontology, a function describes methods, and the process ontology describes a series of steps similar to a workflow. In the WDO approach, abstract workflows are created dynamically, which allows them to adapt to new concepts and relationships.

The Transparent Access to Multiple Biological Information Sources (TAMBIS) [1] is a bioinformatics ontology whose design is based on description logics in order to allow dynamic creation and reasoning about the concepts. Its organization is based on groupings set by the description logic GRAIL. GRAIL uses a hierarchical composition of the concepts (a group of individuals that share common characteristics), the individuals that make up these classes, and bidirectional roles between individuals. TAMBIS uses reasoning services and sanctions to construct new concepts and compose new

concept descriptions. The TAMBIS ontology recognized the importance of distinguishing between various representations of a concept; therefore, it is organized into multilayer divisions. For example, in the bioinformatics world, a structure can be separated into its physical and abstract representations. Thus, the *Generalized Structure* division for a concept is separated into *Physical Structure* and *Abstract Structure*. Also, the ontology has separate concept divisions for biological processes and biological functions. This notion of distinguishing between the possible representations of a concept helps reinforce the idea that separating concepts into categorizations is beneficial. In TAMBIS, types of roles are distinguished through attribute categories. For example, there is the *FunctionalAttribute* category that has relationships between either processes or functions, or between physical and abstract things. Having different types of roles makes relationships between concepts more informative, and it stresses the importance of typecasting roles. Like TAMBIS, the WDO approach adopts the separation of concerns with respect to concepts.

The Semantic Web for Earth and Environmental Terminology (SWEET) ontologies [15] were developed to capture knowledge about Earth System science. A group of scientists have captured thousands of Earth System science terms using the OWL ontology language. There are two main types of ontologies in SWEET: facet and unifier ontologies. Facet ontologies deal with a particular area of Earth System science (earth realm, non-living substances, living substances, physical processes, physical properties, units, time, space, numerics, and data). Unifier ontologies were created to piece together and create relationships that exist among the facet ontologies. Facet ontologies use a hierarchical methodology in which children are specializations of their parent nodes. The SWEET ontologies are currently being used in the Geoscience Network (GEON) [3] as a base vocabulary for most of the ontologies currently available through GEON.

Summary

The workflow-driven ontology was devised to support scientists' ability to capture discipline-specific knowledge that supports their research. Such an ontology focuses on the capture of processes as well as data and reduces the dependence on a technologist to construct an ontology. Workflow-driven ontologies are distinguished from discipline-based ontologies that capture basic knowledge about a discipline by capturing concepts and relationships that are tied to how results are generated. In particular, all defined methods are tied to the inputs, outputs, and other computation-associated relationships required to generate a result from a specified method. The gravity data

ontology is the first comprehensive ontology that was developed using this approach.

The work reported in this paper has transitioned to the development of a prototype WDO API [10] to facilitate the integration and reuse of WDOs by the WDO-It! Tool and other WDO-related tools that are being prototyped. The WDO API is built on top of the Jena2 Ontology API [6] that provides functionality to access OWL ontologies through Java programming. The WDO API offers specific methods that facilitate the development of WDOs, as well as functionality to create abstract workflow specifications. The WDO-It! tool provides a GUI to assist scientists to create new WDOs. Work is in progress to extend domain ontologies into WDOs and to transform abstract workflows to executable workflows.

References

1. Baker, P.G., C.A. Goble, C.A., S. Bechhofer, N.W. Paton, R. Stevens R, and A. Brass, "An Ontology for Bioinformatics Applications," *Bioinformatics*, 15(6):510-520, 1999.
2. Getting Started: Using and Understanding Gravity Data http://paces.geo.utep.edu/grav_database/grav_db_getstart.shtml, March 22, 2004.
3. GEON, The Geosciences Network: Building Cyberinfrastructure for the Geosciences, http://www.geongrid.org/, June 2007.
4. Hinze, W. J., C. Aiken, J. Brozena, B. Coakley, D. Dater, G. Flanagan, R. Forsberg, T. Hildenbrand, G. R. Keller, J. Kellogg, R. Kucks, X. Li, A. Mainville, R. Morin, M. Pilkington, D. Plouff, D. Ravat, D. Roman, J. Urrutia-Fucugauchi, M. Véronneau, M. Webring, and D. Winester, " New standards for reducing gravity data: The North American gravity database," *GEOPHYSICS*, 70: 325-332, 2005.
5. Holom, D. I. and J. S. Oldow, "Gravity reduction spreadsheet to calculate the Bouguer anomaly using standardized methods and constants," Geosphere, 3(2):86–90; doi: 10.1130/GES00060.1, 2007.
6. Jena2 Ontology API, http://jena.sourceforge.net/ontology/index.html, July, 2006.
7. Noy, N. F. and D. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology". *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05*, March 2001.
8. Pinheiro da Silva, P. et al. "Knowledge Provenance Infrastructure," *IEEE Data Engineering Bulletin*, 26(4), pp. 26-32, December 2003.
9. Pinheiro da Silva, P. and A. Gates, CI-Miner, http://trust.utep.edu/ciminer/, 2006.
10. Salayandia, L., P. Pinheiro da Silva, A. Gates, and F. Salcedo, "Workflow-Driven Ontologies: An Earth Sciences Case Study," in *Proceedings e-Science 2006,* Amsterdam, Netherlands, December 2006.
11. Salayandia, L., P. Pinheiro da Silva, A. Gates, and A. Rebellon, "Domain-Level Workflows for Scientific Applications," in *Proceedings 6th OOPSLA Workshop on Domain-Specific Modeling,* October 2006.
12. Salcedo, F., "A Method for Designing Computation-Driven Ontologies in the Geosciences," Master's Thesis, University of Texas at El Paso, May 2006.

13. Smith, B., J. Williams, and S. Schulze-Kremer, "The Ontology of the Gene Ontology," *Proc. AMIA Symp. 2003,* 2003, pp. 609-613.
14. Smith, M. K., C. Welty and D. L. McGuiness, "OWL Web Ontology Language Guide," World Wide Web Consortium (W3C) recommendation, February 2004, http://www.w3.org/TR/owl-guide/.
15. SWEET, "Guide to SWEET Ontologies," http://sweet.jpl.nasa.gov/guide.doc, June 2007.
16. Wessel, P., and W. H. F. Smith, "New version of Generic Mapping Tools released," EOS, *Transactions American Geophysical Union*, 76:329, 1995.