ODISEES: Ontology-driven Interactive Search Environment for Earth Sciences

Matthew T. Rutherford[1], **Elisabeth B. Huffer**[2], **John M. Kusterer**[3], **Brandi M. Quam**[4]
[1,2] Lingua Logica LLC, NASA Langley, Atmospheric Science Data Center, Hampton, VA, United States
[3,4] Science Directorate, NASA Langley, Atmospheric Science Data Center, Hampton, VA, United States

Abstract— *This paper discusses the Ontology-driven Interactive Search Environment for Earth Sciences (ODISEES) project currently being developed to aid researchers attempting to find usable data among an overabundance of closely related data. ODISEES' ontological structure relies on a modular, adaptable concept modeling approach, which allows the domain to be modeled more or less as it is without worrying about terminology or external requirements. In the model, variables are individually assigned semantic content based on the characteristics of the measurements they represent, allowing intuitive discovery and comparison of data without requiring the user to sift through large numbers of data sets and variables to find the desired information.*

Keywords: ontology, semantics, data discovery, search

1. Introduction

Over the past few decades, researchers have often been faced with a unique and very modern problem: the overabundance of usable, relevant data. The exponentially shrinking cost and size of computer hardware components has enabled the storage of vast quantities of data, and the emergence of the Internet has enabled near-instantaneous dissemination of this data. As a result, finding precisely the right data can be like searching for the proverbial needle in a haystack. Oftentimes, researchers must sift through large volumes of closely related information in order to uncover the desired usable information buried there. This is, in many ways, the opposite of what researchers have had to deal with in the past where data was generally much less abundant and more difficult to assess. As such, the scientific and information technology communities face the ever-growing challenge of storing, managing and distributing vast amounts of data and providing user-friendly tools to the scientific communities that hope to use the information therein. The Ontology-Driven Interactive Search Environment for Earth Sciences (ODISEES) project seeks to offer an alternative to traditional methods of storage and organization.

There is a wealth of Earth science data—atmospheric, geological, meteorological, hydrological, etc.—that has grown rapdily over the past few decades. These data are produced through a variety of collection methods and technologies and interpreted by scientists and researchers from a wide, heterogeneous set of disciplines for a similarly large and varied set of purposes. Earth science researchers are often faced with two significant, recurring challenges:

1) Finding data products that are immediately relevant to their research
2) Quickly noticing and understanding the similarities and differences among closely related products and assessing their suitability for a particular purpose

As one of the Distributed Active Archive Centers (DAACs) under the umbrella of NASA's Earth Observing System Data and Information System (EOSDIS), the Atmospheric Science Data Center (ASDC) at the NASA Langley Research Center is a steward of large amounts of Earth science data. The ASDC's purpose is to make its petabytes of data holdings easily available to the public, serving a variety of users [1], including scientists and researchers as well as the general public. Given that scientists and laymen alike are actively accessing and using its archives, the ASDC is met with the unique challenge of catering to a user base with inconsistent knowledge of its data holdings.

The use of ontologies and semantics has emerged as one of many solutions to address these issues. Our implementation of this solution, ODISEES, is being developed to provide researchers with tools for discovering and assessing available ASDC data products.

2. Ontology

There are many interpretations of the term "ontology". Long used by philosophers as a conceptual tool for studying the conditions of existence and for classifying that which exists, ontology has more recently been interpreted and implemented by computer and information scientists—primarily as a computer-readable artifact that can represent the entities and relationships in a domain of interest.

Historically, ontology-based solutions have often started out ambitiously. Researchers in Artificial Intelligence, for instance, once looked to ontology and deductive reasoning systems in an attempt to create truly intelligent systems that could overcome the limitations of expert systems. The Cyc project [2], [3] is one such project, designed to represent the contextual common sense knowledge that humans take for granted when they engage in deductive reasoning. However, backlash over the years against AI in general, deductive reasoning systems like Cyc in particular [4], [5], [6], and top

down ontologies, which try to impose context-independent structure, resulted in less ambitious use cases for ontology. At the same time, the Semantic Web was developing and ontologies emerged as a means for controlling vocabularies and bolstering the exchange of information on the Internet [7]. The Resource Description Framework (RDF) [7], [8], which effectively merged ontology and XML, was developed. Since then, the word "ontology" has largely been used to refer to controlled vocabularies that are used to provide semantic content for marking up data objects for the Internet. In this context, the reasoning and inference capabilities that were critical to AI applications became less important.

While less ambitious uses of ontology have become more widely accepted, the term "ontology" continues to be subject to multiple interpretations. One of the more popular definitions of ontology in an information and computer science context comes from Tom Gruber [9]: "...an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members)." Some liberal interpretations can include taxonomies, relational models, and XML, but these do not provide any semantics. At the other end of this interpretive spectrum are full-fledged logical theories of domain that rely on first-order, second-order, and even modal logic. In developing ODISEES, we were interested in the latter, more robust interpretation and application of ontology: artifacts that model a domain with a good deal of precision and leverage the inferential powers of first and second-order logic.

3. Methodology

ODISEES relies on the ontological classification of measured phenomena represented in ASDC data products, using the resulting characterization of the data to enable effective, expedited discovery and comparison of said data. It was designed to perform three primary tasks in service to data search and discovery:

1) Model the set of objects and concepts that make up the Earth science domain and the relationships among them in order to provide a common domain model to impart meaning to the terms used to describe the domain
2) Identify and define the terms used by specialized user groups within the Earth science community, mapping these terms to the common domain model and creating computer-readable definitions of community-based terms
3) Be usable by humans, databases, and applications that need to interpret model to impart meaning to the terms used to describe the domain

The ODISEES search system is comprised of open-source, commercial-off-the-shelf (COTS), and custom software components that interact with an Earth science ontology and metadata repository to offer long-term solutions to the problem of data discovery in the era of Big Data. The ODISEES ontology and metadata repository provides an ontological and lexical framework for describing ASDC data holdings, as well as support for deductive reasoning and querying capabilities. The aforementioned RDF format was chosen to represent this information. An intuitive web-based user interface was developed, allowing users to quickly sort through and analyze relevant portions of the ontology in order to locate the desired data.

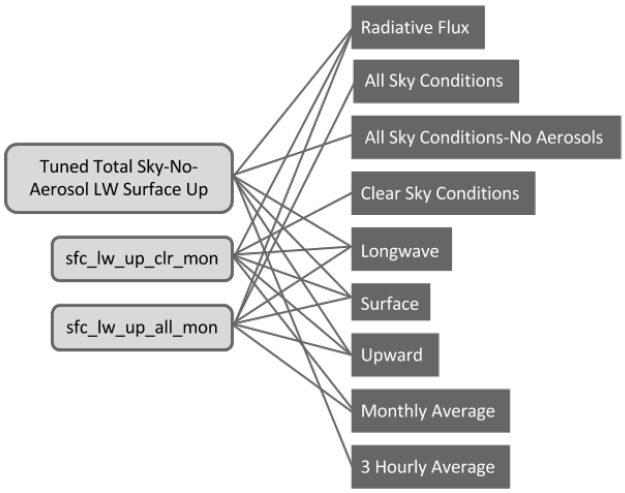


Fig. 1: Variables' semantic content. Variable names are on the left; semantic content is on the right. These attributes were ascertained by examining actual variables and data sets instead of requirements.

3.1 Concept Modeling

ODISEES was, in large part, developed around the idea that recognition is generally faster, easier, and simpler than recall. As such, the model itself is not overly concerned with terminology or nomenclature, and instead puts greater emphasis on the semantics and relationships of concepts. The ontology was structured with the intention of merging a controlled vocabulary with useful deductive reasoning systems that can support semi-intelligent applications. Controlled vocabularies can, however, be overly restrictive and cumbersome for some types of applications, such as text-based searches, insofar as they often require significant effort on the user's part to either memorize a lot of terms or spend a lot of time looking up the correct term. Instead of using this kind of terminology-centric approach, the ODISEES ontology focuses on modeling the domain concepts that give meaning to terms, and treats these terms as first-class objects that refer to the concepts in the domain.

For example, a data variable is described in the ontology as a set of relationships between it and other domain objects. A Radiative Flux variable will have a relationship to a certain

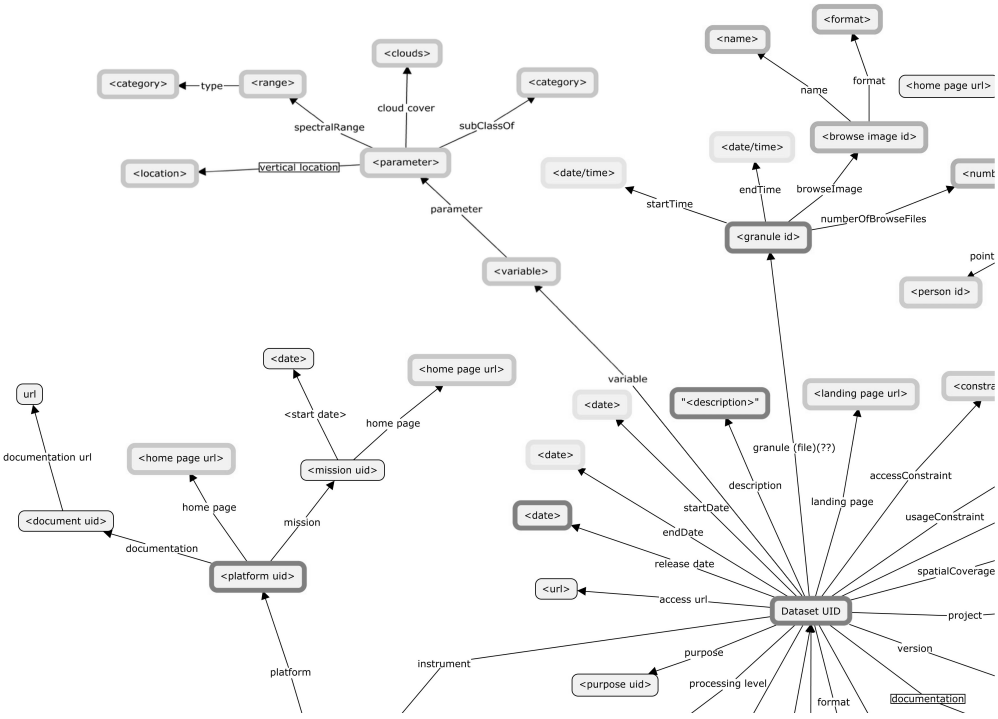


Fig. 2: Data set and variable class models. The data set model is at the bottom right, and the variable model at the top left. Note the line connecting the two models. This indicates that the variable class is an attribute of the data set class.

wavelength or spectral range. The relationship is represented as a binary function that takes a particular atmospheric radiation variable, such as "CERES SW TOA Flux Upwards", and a particular spectral range, such as "shortwave", as arguments. Each variable's set of attributes is determined by analyzing the variable itself, the data sets in which it is included, metadata about the data set, and reviewing product documentation as well as consulting with domain experts. Taken together, the set of relationships describe the variable in a machine-readable format with sufficient detail to allow a scientist to assess, with reasonable precision, the essential characteristics of the observation or model output represented. This representation provides maximum flexibility and extensibility in describing Earth Science data and model outputs because, at any point, new domain objects can be introduced, new relations can be defined, and new relationships among variables and other domain objects can be asserted, without requiring any changes to the underlying data model. As a result, variables, or indeed any object in the model, can be identified and evaluated strictly in terms of their defining attributes and without regard for the naming convention that may have been followed in labeling it. Figure 1 provides an illustration of how this semantic content is attached to variable names.

Each attribute of a given object in the ODISEES model is itself an object and has a label or name. Variable objects are themselves attributes of the data set objects. Figure 2 shows the concept models for both the variable class and the data set class.

A primary strength of this kind of model is that even if the application requirements change, the model can be easily adapted to accommodate these changes. For the ODISEES search application, we want some of the attributes of a data set to be inherited by the variables within that data set. Similarly, if characteristics of the remote sensing instrument used to create the data set have implications for the data it produces, we want the data set or the variables within it to inherit whatever attributes are implied. For example, if a sensor is calibrated to measure radiation in a particular spectral range, the spectral range associated with a data variable produced by that instrument can be automatically inferred from the fact that it was produced by that instrument. The inferential capabilities of our logic-based model were used to materialize many of the assertions we wanted to drive the application.

3.2 Variable Discovery and Comparison

As mentioned in the beginning of Section 3, the core purpose of ODISEES is not just to represent or conceptualize information, but rather to make information more discoverable and thereby usable. More specifically, ODISEES is designed to leverage the concept modeling system described in Section 3.1 to enable users to more easily and effectively discover and compare variables from many distinct data sets. To this end, ODISEES uses semantic content attached to

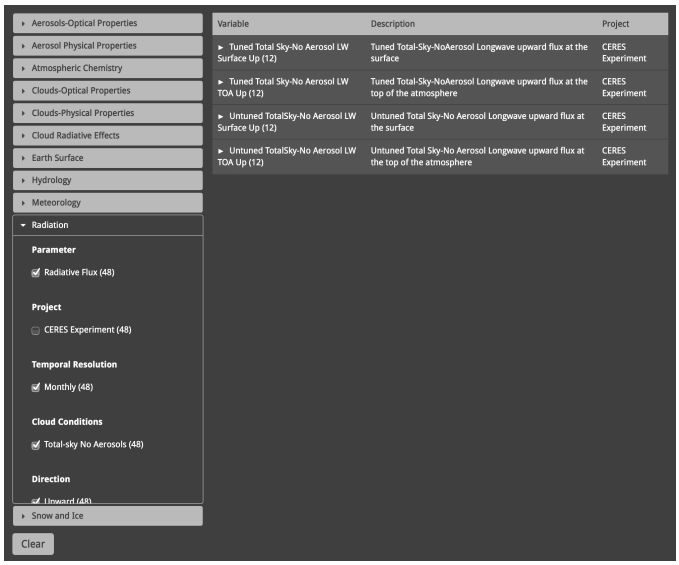
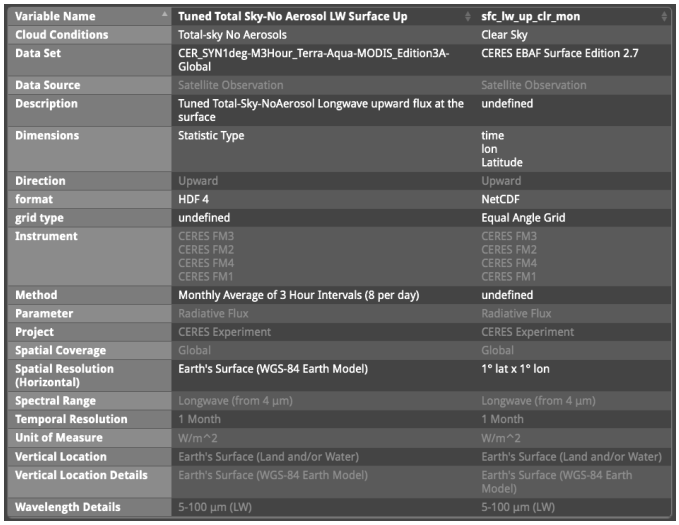


Fig. 3: Attribute filter selection with the ODISEES user interface.



Variable Name	Tuned Total Sky-No Aerosol LW Surface Up	sfc_lw_up_clr_mon
Cloud Conditions	Total-sky No Aerosols	Clear Sky
Data Set	CER_SYN1deg-M3Hour_Terra-Aqua-MODIS_Edition3A-Global	CERES EBAF Surface Edition 2.7
Data Source	Satellite Observation	Satellite Observation
Description	Tuned Total-Sky-NoAerosol Longwave upward flux at the surface	undefined
Dimensions	Statistic Type	time lon Latitude
Direction	Upward	Upward
format	HDF 4	NetCDF
grid type	undefined	Equal Angle Grid
Instrument	CERES FM3 CERES FM2 CERES FM4 CERES FM1	CERES FM3 CERES FM2 CERES FM4 CERES FM1
Method	Monthly Average of 3 Hour Intervals (8 per day)	undefined
Parameter	Radiative Flux	Radiative Flux
Project	CERES Experiment	CERES Experiment
Spatial Coverage	Global	Global
Spatial Resolution (Horizontal)	Earth's Surface (WGS-84 Earth Model)	1° lat x 1° lon
Spectral Range	Longwave (from 4 µm)	Longwave (from 4 µm)
Temporal Resolution	1 Month	1 Month
Unit of Measure	W/m^2	W/m^2
Vertical Location	Earth's Surface (Land and/or Water)	Earth's Surface (Land and/or Water)
Vertical Location Details	Earth's Surface (WGS-84 Earth Model)	Earth's Surface (WGS-84 Earth Model)
Wavelength Details	5-100 µm (LW)	5-100 µm (LW)

Fig. 4: Variable comparison with the ODISEES user interface.

individual variables. This content is encoded as a set of RDF assertions that uniquely describes each variable. We say "uniquely" because, even though variables may have many attributes in common, the complete set of assertions that describes any given individual variable is unique to that variable.

The set of RDF assertions that describe the data variables are used as filters that let users specify criteria and thereby narrow their search to a set of results that satisfy all and only those criteria. Figure 3 shows a set of filters and the variables which satisfy the selected options as they appear in the ODISEES user interface.

Once the desired variables are selected, users can generate a comparison table, which displays each variable's respective semantic content items and highlights differences between them to aid and inform the comparison process. Figure 4 shows this comparison feature in the ODISES user interface.

3.3 Results

The ODISEES beta version is currently deployed and maintained at the ASDC, and is accessible to the public at odisees.larc.nasa.gov. We are actively developing improvements to the search capabilities and increasing the amount of searchable data in the ODISEES repository. There are currently 91 data sets represented in the ontology, and new ones are being added regularly. The ontology is still relatively small—400,931 RDF triples—but it's expected to grow significantly over the next two years. Initial user feedback has been generally positive, with many users attesting to the present and future usefulness of the tool. The test user group is composed of developers and researchers from NASA, NOAA, the EPA, multiple universities, and several other organizations.

4. Discussion

The ODISEES project discussed in this paper presents an adaptable, modular solution to some challenges posed by the extreme abundance of closely related data. Furthermore, it demonstrates the potential usefulness of ontology-based applications in solving issues posed by the growing overabundance of data.

Several features, including text search, simple data subsetting with the Open-source Project for a Network Data Access Protocol (OPeNDAP), and various improvements to the web interface, are planned for future releases. Additionally, development has begun on the Ontology-based Metadata Portal for Unified Semantics (OlyMPUS) [10], a metadata ingest system which leverages the same ontological structure as ODISEES. The ODISEES-OlyMPUS end-to-end system will support both data consumers and data providers, enabling the latter to register their data sets and provision them with the semantically rich metadata that drives ODISEES' data discovery and access service for data consumers.

5. Acknowledgments

The authors would like to thank the Atmospheric Data Center at NASA's Langley Research Center for their ongoing support.

References

[1] ASDC, "About the Atmospheric Science Data Center," 2015. [Online]. Available: https://eosweb.larc.nasa.gov/more-about-asdc.
[2] Cyc, "Cyc Homepage," 2015. [Online]. Available: http://www.cyc.com/.
[3] D. Lenat, S. Laningham, "Doug Lenat on Cyc, a Truly Semantic Web and Artificial Intelligence AI," 2008. [Online]. Available: http://www.ibm.com/developerworks/podcast/dwi/cm-int091608txt.html.

[4] J. Friedman, "The sole contender for ai" in *Harvard Science Review*, 2003.

[5] D. Lenat, G. Miller, and T. Yokoi, "Cyc, wondernet, and edr: critiques and responses" in *Communications of the ACM*, vol. 38, no. 11, 2003.

[6] J. Taubarer, "AI Founder Blasts Modern Research," 2003. [Online]. Available: http://archive.wired.com/science/discoveries/news/2003/05/58714?currentPage=all.

[7] D. Allemang, J. Hendler, *Semantic Web for the Working Ontologist*. Elsevier, 2011.

[8] J. Taubarer, "What is RDF?," 2006. [Online]. Available: http://www.xml.com/pub/a/2001/01/24/rdf.html.

[9] T. Gruber, "Ontology" in *Encyclopedia of Database Systems*, 2009.

[10] J. Gleason, E. Huffer, A. Ross, P. McInerey, P. Mehrotra, P. Rinsland, "Ontology-based metadata portal for unified semantics (olympus)", Proposal submitted in response to National Aeronautics and Space Administration (NASA) Research Announcement (NRA) Advanced Information Systems Technology Program, 2014.