An Ontology Mapping Approach to Integrating Earth Science Metadata

Richard M. Keller[1], Rajkumar Thirumalainambi[2], Joseph C. Coughlan[1]

Intelligent Systems Division[1]
Perot Systems Government Services[2]
NASA Ames Research Center
Mail Stop 269-2, Moffett Field, CA 94035-1000
Richard.Keller@nasa.gov, rajkumar@mail.arc.nasa.gov, Joseph.C.Coughlan@nasa.gov

Abstract

One of the main barriers to exploiting the great wealth of global earth science data available today is that researchers are unable to rapidly search and find data relevant to their studies. This data is spread across a large number of archives maintained by different institutions employing a bewildering array of different data description languages. In this paper, we describe a metadata federation approach designed to support queries across multiple earth science data archives without requiring the adoption of a unified metadata standard. Our ontology-based approach employs a central metadata transformation facility capable of integrating heterogeneous metadata using a set of translators and wrappers. This shifts the burden of federation from the data provider to the central metadata facility, acknowledging that not all data providers have the motivation or resources to comply with externally-imposed metadata standards. We demonstrate the feasibility of this approach with a proof-of-concept prototype that federates metadata across two earth science data archives – one containing NASA data and the other containing USDA data – despite the differences in their metadata languages.

Introduction

NASA's earth science missions collect data that must often be used in combination with non-NASA data to answer complex science questions. Earth scientists must leverage data collected by NASA, as well as by other government agencies, public and private sector organizations, and foreign governments to accomplish their science objectives. One of the main barriers to using global earth science data effectively is that researchers are unable to rapidly search and find relevant data across multiple archives with heterogeneous formats from different institutions. Each on-line archive has its own search mechanism and its own set of metadata keywords and values for describing stored datasets. As a result, finding data is a time-consuming exercise.

To address the problem of locating earth science data, NASA has developed the EOS Data Gateway [1] (EDG), which provides search services that range over NASA's nine Distributed Active Archive Centers (DAACs) [2] containing NASA-funded data. Unfortunately, most metadata federation schemes – including the scheme employed by the EDG – rely on the application of standards that place significant burdens on the data providers, who must modify their metadata management approach to be compliant. In particular, all of the DAACs are required to publish their metadata in a common format established by NASA. While it may be possible for NASA to mandate standards unilaterally for the projects it funds, this approach does not support integration with valuable non-NASA earth science datasets based upon on divergent metadata schemes.

In this paper, we describe a metadata federation approach that allows each data archive to publish its own metadata format without the burden of adopting an externally-imposed standard. This approach employs a central metadata transformation facility capable of integrating NASA metadata with metadata from external data sources using a set of translators and wrappers. This shifts the burden of federation from the data provider to the central metadata facility, acknowledging that not all providers of useful data have the motivation or resources to comply with NASA metadata standards.

As a proof-of-concept, we are applying this approach to federate metadata from NASA's Biogeochemical Dynamics data archive (BGC-DAAC) with metadata describing the U.S. Department of Agriculture's (USDA) STEWARDS watershed data archive. The BGC-DAAC [3] holds most of the NASA's terrestrial field experimental data, constituting 20% of all EOS products. The USDA's Agricultural Research Service (ARS) has been collecting soil, water quality, and climate data in benchmark watersheds for almost a century, and has recently built the STEWARDS archive to hold the data [4]. The STEWARDS data is valuable to NASA researchers who can use the data to validate biogeochemical models and remote sensing observations; the NASA BGC-DAAC data is valuable to USDA researchers who require additional water and land cover data to improve the fidelity of their conservation effects assessment models.

Metadata Integration Framework

The challenge of metadata integration across multiple sources is to provide a scalable, maintainable approach without requiring data providers to alter their metadata curation or publishing methods. Our SemanticIntegrator framework [5] uses semantic integration techniques [6] to federate across a distributed set of data archives. SemanticIntegrator was initially designed to integrate various heterogeneous sources of planetary exploration data, including field-collected geology data, satellite imagery, GIS data, and physical/optical properties of minerals. To support integration, the framework requires that an ontology be developed to describe the metadata for each of the source data archives (SDAs) to be integrated. In addition, an overarching ontology must be developed that incorporates a comprehensive set of metadata distinctions found across the various sources. Conceptually, this ontology specifies a *virtual* data archive (VDA) that unites metadata across all of the SDAs.

Each data archive ontology consists of:

- a set of entities pertinent to the archives (e.g., *fieldSites*, *instruments*, *measurements*, *PIs*, *dataGranules*, *projects*);
- a set of attributes defined for each archive entity (e.g., *sensorType*, *collectionDate*, *collectionMethod*, *dataUnits*, *PIcontactInformation*);
- a set of semantic relationships that establish cross-linkages among the archive entities (e.g., PI *leads* project, measurement *collectedBy* instrument, measurement *collectedAt* fieldSite); and
- a set of logical axioms that allow for specific forms of automated reasoning, classification, and constraint maintenance over the archive entities, attributes, and relationships (e.g., "*all fieldWorkSites are dataCollectionSites*", "*measurements must be collectedAt dataCollectionSites*", "*only PIs can lead projects*").

In essence, ontologies provide a vocabulary for describing data, constraints on data, and implications of the data.

Generally, the metadata languages defined for the VDA and the individual SDAs will exhibit significant differences. At a minimum, different terms ("dataset", "collection", "fileset") will be used to express the same metadata concept across different archives. To accommodate these differences, translation is necessary to formulate a given query in a manner that is compatible with the different archives. We use the VDA ontology as a mediating language through which all queries must pass; a query can be posed using the metadata query language for a given SDA and then translated into the query languages for all other SDAs by passing through the VDA. A set of translation rules specifies how to translate to and from the VDA language. These translation rules are interpreted and executed by a Data Source Mediator (DSM), the central module of SemanticIntegrator. Wrappers are created to receive translated ontology language queries and resolve these into actual queries appropriate to the native SDA query API (e.g., web services, HTTP request, database connectors, or other specialized APIs). A summary of this metadata integration framework is shown in Figure 1.

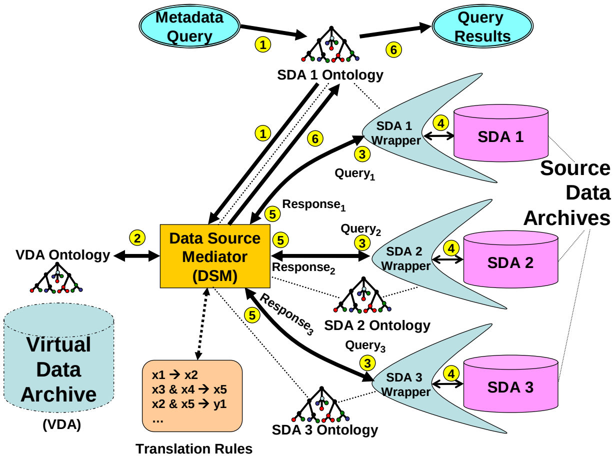


Figure 1 – Metadata Integration Framework: (1) Metadata query is issued using terms from SDA#1 ontology and is transmitted to DSM, which uses translation rules to reformat query in terms of the VDA ontology (2). DSM translation rules next rewrite VDA query in terms of SDA#2 and SDA#3 ontologies, and queries for all three SDAs are sent to their respective wrappers (3). Finally wrappers query the SDAs archives to retrieve metadata (4) and transmit results back through DSM (5), which translates results back into the SDA#1 ontology language (6).

A key feature of this approach is that the individual SDA requires no modification to participate in federation by the VDA. All that is required is a means of programmatically querying the SDA. The wrapper, translation rules, and ontology can be developed in a manner independent of the SDA's internal workings and without the commitment of dedicated SDA resources and personnel, if necessary.

To simplify translation among the VDA and SDA metadata languages, they are each expressed in RDF [7], which will serve as a *lingua franca* for metadata interchange. RDF is a simple ontology language for expressing metadata relationships using triples of the form <*subject predicate object*>. For example, one might express the fact that a dataset contains temperature data collected in 1993 from Wilbur, Kansas in RDF as follows:

<dataset1 containsData tempData1> &
<tempData1 measures temperature> &
<tempData1 collectionYear 1993> &
<tempData1 collectionPlace WilburKansas>

The next section describes our application of this metadata integration framework as part of building a unified search engine across two heterogeneous Earth Science datasets.

Application of Framework to Integrating USDA and NASA Metadata

We applied the metadata integration framework to search metadata across two SDAs exhibiting very different structural characteristics – *STEWARDS* and *BGC-DAAC*:

- *USDA/ARS STEWARDS*: The STEWARDS (Sustaining the Earth's Watersheds Agricultural Research Data System) archive [4] is being developed by a team of USDA ARS researchers to aggregate information on climate, water, and soil, as well as on management and economic practices for 14 benchmark U.S. watersheds participating in the Conservation Effects Assessment Project (CEAP) [8]. Data for these watersheds have been collected from the early 20th century onward in varying formats. Each watershed measures, collects, and records data using differing procedures and employs slightly different terminology; this makes it very challenging to compare data across watersheds and generate meaningful analyses. The watershed data, which have never been centralized, are being collected and stored as part of the STEWARDS project to prepare the way for performing data integration and cross-watershed evaluation studies. The data and metadata are stored in relational database tables and are presented using ArcInfo [9], a web-based GIS interface.

- *NASA BGC-DAAC*: This archive, managed by the Oak Ridge National Laboratory (ORNL), is the primary source for NASA biogeochemical and ecological data and models useful for environmental research. Data held in the BGC-DAAC have been collected and archived from field observations, aircraft surveys, satellite collection, and from computer model output. These data are typically stored in investigator-supplied files that are maintained on the ORNL DAAC servers. The BGC-DAAC metadata is harvested from various descriptive files uploaded by science project personnel and stored in a relational database. The XML-formatted metadata contain information about the research being performed, the sensors used, the parameters measured, the data collection time and place, as well as recording a set of science keywords characterizing the data [10].

Applying the SemanticIntegrator framework to integrate metadata across these two archives involved the following:

1. Develop ontologies to represent the STEWARDS and BGC-DAAC archives.
2. Develop a cross-cutting ontology (the VDA ontology) to serve as an interlingua for transformations between the STEWARDS and BGC-DAAC archives.
3. Engineer a set of translation rules to convert statements (i.e., RDF triples) in the STEWARDS or BGC-DAAC ontologies into statements in the VDA ontology, and vice versa.
4. Implement wrappers to convert ontology queries expressed in terms of either the STEWARDS or BGC-DAAC ontologies into native metadata queries that can be posed against the actual archive metadata; a reverse wrapper must also be implemented to convert retrieved metadata results back to a set of RDF triples.

VDA Ontology

The VDA Ontology (Figure 2) captures common concepts that are important to describing both the STEWARDS and BGC-DAAC data. For this proof-of-concept demonstration, we generalized beyond the native metadata representations present in either STEWARDS or BGC-DAAC when designing the VDA Ontology, but due to resource limitations, we did not put a great deal of effort into creating a comprehensive ontology. The VDA Ontology incorporates various types of scientific measurements, scientific units, spatial and temporal extents, geopolitical representations, and other concepts related to data collection and storage, including the social structures that support these activities. Although we did not base the VDA Ontology on any existing ontologies, incorporation of concepts from SWEET [11] would be a logical option to consider going forward.

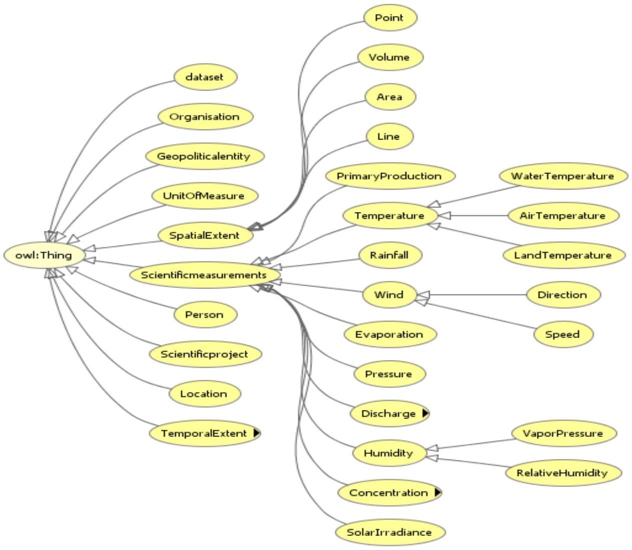


Figure 2 – Virtual Data Archive (VDA) Ontology

BGC-DAAC Ontology

We derived the BGC-DAAC Ontology (Figure 3) directly from the metadata language used to describe the data. We included only concepts, attributes, and relations directly analogous to the XML elements specified in the BGC-DAAC metadata language. (The metadata language specification was supplied to us by ORNL, and is based on metadata standards published by the Federal Geographic Data Committee (FGDC) [12].) The ontology incorporates project and contact information, keywords that cover the general nature of the scientific measurements collected, information about where and when the data were collected, citations to publications describing the data collection activity, and other information (See Figure 3).

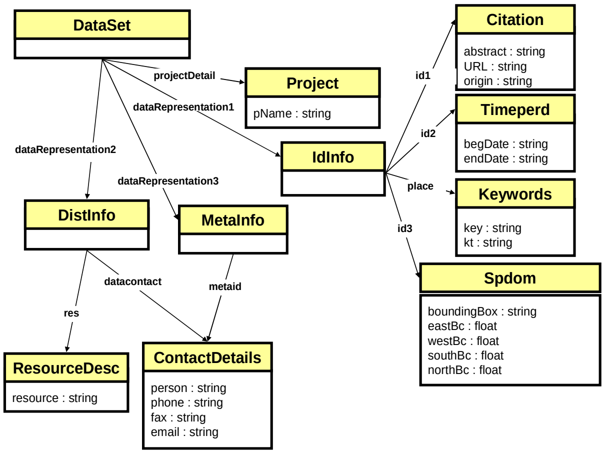


Figure 3 – Ontology for BGC-DAAC Data Archive

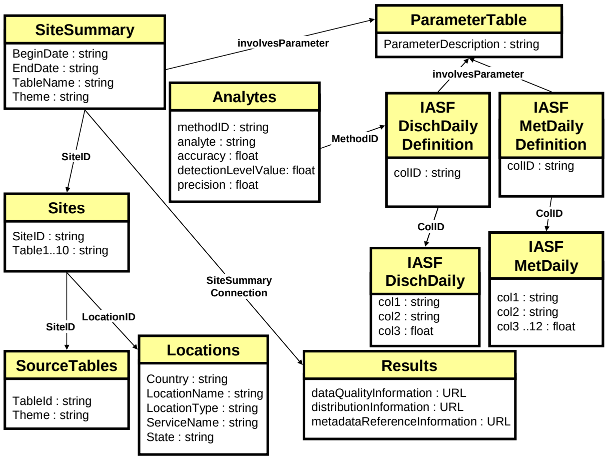


Figure 4 – Ontology for STEWARDS Data Archive

STEWARDS Ontology

Unlike the BGC-DAAC archives, STEWARDS was not designed to use a distinct metadata language. Both its metadata and data are being stored in tables within a relational database. These tables will eventually hold information covering 14 watersheds over a 20-year timespan. The STEWARDS metadata is much more fine-grained than the BGC-DAAC metadata, and includes comprehensive descriptions of all measurements as well the data collection methodology. To create the BGC-DAAC Ontology (Figure 4), we imported the relational database schema provided by the USDA into RDF using the Protégé DataMaster plugin [13]. The import procedure automatically converts relational database tables into classes and table columns into slots. (Columns that are foreign keys get converted into links between classes.)

Translation Rules

The mapping between SDA and VDA ontologies is performed via a set of ontology translation rules. These rules are executed by the Data Source Mediator (DSM – Figure 1), which utilizes the Jess rule engine [14] to support inferencing and query rewriting. The DSM also serves as a registry that maintains information about the different data archives being integrated. Translation rules are simple if-then rules, where each clause takes the form of an RDF triple. Following are two simple translation rules:

R1: if (?Project1 daac:pName ?Name1)
then (?Project1 vda:projectName ?Name1)

R2: if (?Loc1 stewards:locationName ?Name2)
then (?Loc1 vda:projectName ?Name2)

Each of these rules derives the value of the projectName property, which is defined in the VDA ontology. R1 infers this value from the pName property of a project instance in the BGC-DAAC ontology, whereas R2 infers the same property based on the locationName of a location instance in the STEWARDS ontology. These two rules highlight some differences in how data storage is conceptualized in these different archives. Although the BGC-DAAC archive includes the notion of a project, the STEWARDS archive does not. The closest to the 'project' concept in the STEWARDS archive is the notion of a watershed 'location'. For each watershed location, there exist a number of field sites with associated datasets. Similarly, for each project in the BGC-DAAC, investigators carry out measurements at multiple sites, and multiple datasets are associated with each site.

Rules R1 and R2 are invoked via backchaining when the following query is issued:

Q1: (* vda:projectName *)

Q1 infers the vda:projectName property for all vda:ScientificProject instances based on the property values stored in the STEWARDS and BGC-DAAC archives.

As another illustration of the type of ontology translation rules necessary to support this earth science metadata integration task, consider rule R3 below. This rule creates measurement subclasses in the VDA ontology based on the science theme keywords assigned to datasets stored in the BGC-DAAC archive. For each keyword associated with a dataset, a corresponding subclass of measurement is inferred.

```
R3:
if (?Project1 daac:hasDataset ?Dataset1)
   (?Dataset1 daac:dataRepresentation1 ?Idinfo1)
   (?Idinfo1 daac:theme ?Keyword1)
   (?Keyword1 daac:kt ?key1)
then
   (?Keyword1 rdfs:subClassOf vda:ScientificMeasurement)
```

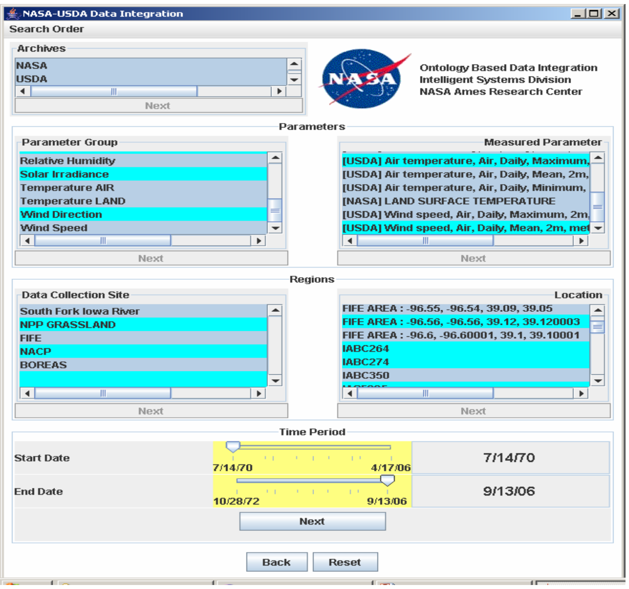
A logical alternative to using our custom rule language plus a Jess-based rule engine would have been to use SPARQL [15]. However, SPARQL was still under development at the time SemanticIntegrator was being developed. In addition, we found the need for sophisticated mapping functions that could be implemented simply using Jess, but might have been more difficult to implement using SPARQL extension functions.

Wrappers

For each source data archive, a separate wrapper is invoked by the DSM to satisfy metadata queries. These queries take the form of RDF triples with wildcards in either the subject or object position (cf. Q1 above). The wrapper translates ontology-based triple queries into native queries against the archives. The native queries may involve web service calls or other types of specialized API calls. ORNL is in the process of developing web services capable of providing access to BGC-DAAC metadata, and the USDA is also planning to develop this capability. As these services are not yet functional, we obtained a portion of the metadata from both archives and stored it on our servers. The STEWARDS data were stored in a Microsoft Access database using a schema identical to the one STEWARDS uses for its data storage. The BGC-DAAC metadata were stored as XML files on our file server. We imported the metadata from STEWARDS and BGC-DAAC into separate Jena [16] repositories and built wrappers to query the metadata instances in those repositories. (Note: We used Jena as a matter of expediency, but our plan is to reimplement these wrappers so that they access the STEWARDS database and BGC-DAAC XML files directly, rather than through Jena. This would be a more in the spirit of the metadata integration framework pictured in Figure 1.) Eventually, when USDA and ORNL publish their web services, the wrappers can be swapped out and replaced with new ones that access the actual metadata archives maintained by USDA and ORNL.

Query Interface

The query interface shown in Figure 5 allows the user to query both the STEWARDS and BGC-DAAC archive metadata using a unified interface. The user makes GUI selections to choose the archives to query, as well as the parameters, the spatial regions of interest, and the time period. After making these selections, the user is presented with a listing of datasets that meet the criteria along with metadata describing the datasets, including information about how to access the data (Figure 6).



Archives
NASA
USDA

Ontology Based Data Integration
Intelligent Systems Division
NASA Ames Research Center

Parameters

Parameter Group
Relative Humidity
Solar Irradiance
Temperature AIR
Temperature LAND
Wind Direction
Wind Speed

Measured Parameter
[USDA] Air temperature, Air, Daily, Maximum,
[USDA] Air temperature, Air, Daily, Mean, 2m,
[USDA] Air temperature, Air, Daily, Minimum,
[NASA] LAND SURFACE TEMPERATURE
[USDA] Wind speed, Air, Daily, Maximum, 2m,
[USDA] Wind speed, Air, Daily, Mean, 2m, met

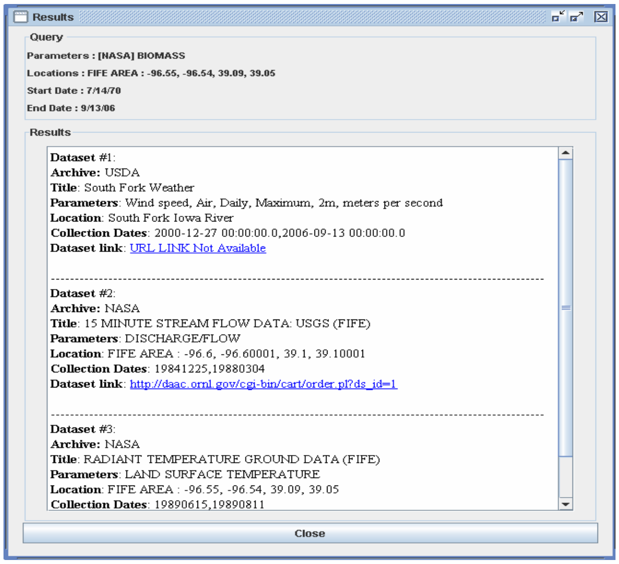
Regions

Data Collection Site
South Fork Iowa River
NPP GRASSLAND
FIFE
NACP
BOREAS

Location
FIFE AREA : -96.55, -96.54, 39.09, 39.05
FIFE AREA : -96.56, -96.56, 39.12, 39.120003
FIFE AREA : -96.6, -96.60001, 39.1, 39.10001
IABC264
IABC274
IABC350

Time Period
Start Date 7/14/70 4/17/06 7/14/70
End Date 10/28/72 9/13/06 9/13/06

Figure 5 – Query interface



Query
Parameters : [NASA] BIOMASS
Locations : FIFE AREA : -96.55, -96.54, 39.09, 39.05
Start Date : 7/14/70
End Date : 9/13/06

Results

Dataset #1:
Archive: USDA
Title: South Fork Weather
Parameters: Wind speed, Air, Daily, Maximum, 2m, meters per second
Location: South Fork Iowa River
Collection Dates: 2000-12-27 00:00:00.0,2006-09-13 00:00:00.0
Dataset link: URL LINK Not Available

Dataset #2:
Archive: NASA
Title: 15 MINUTE STREAM FLOW DATA: USGS (FIFE)
Parameters: DISCHARGE/FLOW
Location: FIFE AREA : -96.6, -96.60001, 39.1, 39.10001
Collection Dates: 19841225,19880304
Dataset link: http://daac.ornl.gov/cgi-bin/cart/order.pl?ds_id=1

Dataset #3:
Archive: NASA
Title: RADIANT TEMPERATURE GROUND DATA (FIFE)
Parameters: LAND SURFACE TEMPERATURE
Location: FIFE AREA : -96.55, -96.54, 39.09, 39.05
Collection Dates: 19890615,19890811

Figure 6 – Results display

What is significant about the interface lies behind the scenes. To populate the interface, the GUI backend formulates a series of STEWARDS ontology queries and forwards each to the DSM. The DSM routes each query directly to the STEWARDS wrapper, which interrogates the STEWARDS archive and returns any matching results. The DSM also translates each STEWARDS query into a VDA ontology query and then into a BGC-DAAC ontology query. This query is passed to the BGC-DAAC wrapper, which interrogates the BGC-DAAC and returns matching results to be combined with the STEWARDS results. All necessary translations are encoded by rules within the DSM, so the USDA application developers are isolated from all details of interacting with the BGC-DAAC metadata language.

Discussion

Our work shares a similar focus with several other projects that fuse heterogeneous scientific datasets and resources using semantic approaches (e.g., GEON [17], SEEK [18], VSTO [19], ISIS [20], AnnoTerra [21]). These projects use various mechanisms to map metadata across data sources for purposes of integration [22]. In some cases, metadata mappings are engineered explicitly by specifying correspondences between data source ontologies (e.g., as in AnnoTerra); in other cases rules are used to specify mappings (e.g., as in SemanticIntegrator, SEEK, ISIS). Our experience is that simple one-to-one term correspondences are insufficient to capture real-world mappings between complex ontologies. The manner in which data is conceptualized and encoded can vary considerably across data sources. A concept found in one ontology may have no direct analog in another, and may need to be inferred based on the presence or absence of other terms. In these cases, more sophisticated mapping mechanisms are required. Rules, on the other hand, are more flexible and powerful. In addition, rules can be reused across domains. However, the use of rules is inherently less efficient because this approach relies on a rule engine to perform mappings. In our prototyping efforts, we have expended considerable effort on implementing caching mechanisms to speed up the rule engine performance and improve system responsiveness.

Conclusion

The scientific data collected during NASA's Earth observation missions and generated by NASA-sponsored earth science research are archived in different data archives across the country. Using metadata integration technology, we hope to enable NASA scientists to submit one query that searches across all NASA earth science data holdings, as well as key non-NASA data holdings – regardless of any differences in metadata standards employed. These benefits are symmetrical; scientists with access to non-NASA data should benefit similarly by improving their ability to query NASA's environmental holdings. This type of improved access to metadata can play a vital role by allowing scientists to discover new datasets to help them further their understanding of the Earth's environmental problems.

Acknowledgments

We would like to thank Bruce Vandenberg from the USDA ARS and Giri Palanisamy from ORNL for providing metadata used as part of this research and for their guidance in interpreting the metadata. We also thank Shawn Wolfe, Dan Berrios, David Hall, and Ian Stuken from NASA for their assistance and their seminal role in the design and development of SemanticIntegrator. We gratefully acknowledge funding from the Earth Science Division of NASA's Science Mission Directorate.

References

[1] EOS Data Gateway http://delenn.gsfc.nasa.gov/~imswww/pub/imswelcome/.
[2] NASA DAAC, http://nasadaacs.eos.nasa.gov/about.html.
[3] BGC DAAC, http://www.daac.ornl.gov/.
[4] STEWARDS, Sustaining the earth's watersheds: STEWARDS Agricultural Research Data System http://arsagsoftware.ars.usda.gov/stewards/.
[5] R. M. Keller, D. C. Berrios, S. R. Wolfe, D. R. Hall, and I. B. Sturken, 2006, "Semantic Integration of Heterogeneous NASA Mission Data Sources", AAAI Fall Symposium: Semantic Web for Collaborative Knowledge Acquisition, Arlington, VA.
[6] N. F. Noy , A. Doan , A. Y. Halevy, 2005, "Semantic integration", AI Magazine, v.26 n.1, p.7-9.
[7] F. Manola and E. Miller (eds), 2004, "RDF Primer", http://www.w3.org/TR/rdf-primer.
[8] CEAP, "Conservation effects assessment project", http://www.nrcs.usda.gov/Technical/nri/ceap/.
[9] ArcInfo, http://www.esri.com/software/arcgis/arcinfo/.
[10] BGC-DAAC Metadata, Oak Ridge National Labs, http://mercury.ornl.gov/metadata/ornldaac/xml/daac/.
[11] SWEET, 2005, "Semantic Web for Earth and Environmental Terminology". http://sweet.jpl.nasa.gov/.
[12] FGDC, "Federal geographic data committee", http://registry.gsdi.org/.
[13] DataMaster, "A plug-in for importing schema structure and data from relational databases into Protégé", http://protegewiki.stanford.edu/index.php/DataMaster.
[14] Jess, "A rule engine for Java platform", http://herzberg.ca.sandia.gov/jess/.
[15] SPARQL, http://www.w3.org/TR/rdf-sparql-query/
[16] B. McBride, 2002, "Jena: A Semantic Web Toolkit", IEEE Internet Computing, v6 no.6, pp. 55-59.
[17] GEON, http://www.geongrid.org/.
[18] S. Bowers & B. Ludaescher "Towards a Generic Framework for Semantic Registration of Scientific Data" Semantic Web Technologies for Searching and Retrieving Scientific Data (SCISW), Sanibel Island, Florida, 2003.
[19] VSTO, http://vsto.hao.ucar.edu/.
[20] Leclercq, E. Benslimane, B. Yetongnon, K., "ISIS: a semantic mediation model and an agent based architecture for GIS interoperability", IDEAS, p. 87, Intnl. Database Engineering & Applications Symposium, 1999.
[21] D. Ramagem, B. Margerin, J. Kendall, "AnnoTerra: Building an Integrated Earth Science Resource Using Semantic Web Technologies," IEEE Intelligent Systems, vol. 19, no. 3, pp. 48-57, 2004.
[22] N. F. Noy, "Semantic Integration: A Survey Of Ontology-Based Approaches". SIGMOD Record, Special Issue on Semantic Integration, 33 (4), December, 2004.