PID-84

Virtual Metadata Catalogs:
Augmenting Metadata Catalogs with Semantic Representations

Yolanda Gil, Varun Ratnakar, and Ewa Deelman
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
gil@isi.edu, varunr@isi.edu, deelman@isi.edu

Abstract

A common approach to managing large, heterogeneous, and distributed collections of data is to separate the data itself (and its physical rendering in replicas) from the *metadata* that describes the nature of the data (often called logical data descriptions). Metadata catalogs store descriptive information (metadata attributes) about logical data items. These catalogs can then be queried to retrieve the particular logical data item that matches the desired criteria. However, the query has to be formulated in terms of the metadata attributes defined for the catalog. Our work explores the notion of *virtual metadata*, where catalogs can be queried using metadata attributes not originally defined in the catalog. We use semantic web standards to query catalogs with virtual metadata and reason about how to map them into existing metadata attributes.

1 Introduction

An integral part of today's large-scale science is the identification and access of large data sets. To support a scalable solution, many systems distinguish between data cataloging and data storage. Data cataloging is designed for ease of publication of data characteristics (metadata attributes) and for ease of querying for data products based on the desired metadata attributes. Having uniquely identified the desired data products (by obtaining an identifier) then enables data access from an appropriate storage location. Metadata attributes and unique identifiers are stored in metadata catalogs, often accessible as services [Singh *et al*., 2003; Myers *et al.*, 2004]. A central goal is the distributed management of data collections that evolve over time and the consumption of those collections by an entire community with diverse conceptualizations of the data and a variety of uses.

We have developed an approach that augments the existing metadata catalogs with semantic representations to create *virtual metadata catalogs*. We augment metadata catalogs with a semantic layer that supports queries in terms of *virtual metadata attributes*, resulting in virtual metadata catalog services. These attributes are virtual in that they are not really used in the implementation of the catalog. However, virtual metadata attributes can be used to query the catalog transparently as if they actually were associated with the data. To support this functionality, the virtual metadata attributes need to be mapped to the metadata attributes that are actually contained in the catalog (*actual attributes*).

2 Virtual Metadata Catalogs

Figure 1 illustrates our approach. By using alternative ontologies we can deliver alternative virtual metadata catalog services that can be built on top of the same underlying metadata catalog. Users may also define their own virtual metadata attributes by creating customized ontologies, effectively creating personalized metadata catalogs. Our implementation of a virtual metadata catalog was developed using the MCS metadata catalog [Singh *et al*., 2003]. We use OWL in combination with rules to express the query, the shared domain ontologies, and the virtual metadata attributes and mappings. The original query is provided as an OWL document that includes references to the domain ontologies from where the virtual metadata attributes in the query are drawn. The query may also reference terms from a generic catalog ontology that we have created. The purpose of this ontology is to define terms such as "files", "views", "collections", that are used in typical queries to MCS. The central component of the architecture is the Query Mapping module. It takes the OWL query and turns it into an MCS query that uses the metadata attributes that actually appear in the catalog. The MCS query is then submitted to MCS, which returns all the references to data stored in it that satisfy the query. For this work we have used data from three different domains: climate modeling, earthquake science, and workflow execution tracking and performance.

The query mapping process is composed of three major steps. First, the basic query constituents are created by running the reasoner and generating attribute/value pairs based on the mapping rules. These virtual metadata attribute value pairs are then converted into metadata attribute/value pairs by selecting the relevant subset of the triples and applying the relevant mapping rules. Another mapping performed in this step is the conversion of the values from the XML Schema Datatypes in the original OWL to the ones that are expected by the MCS database implementation. Finally, the

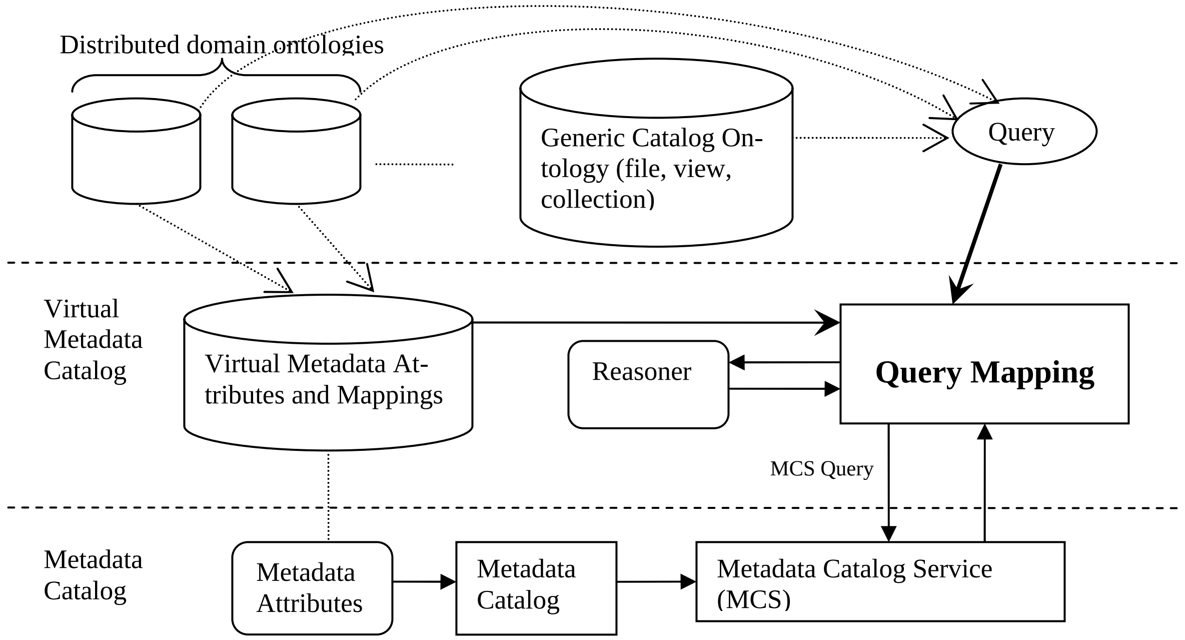
Distributed domain ontologies
Generic Catalog Ontology (file, view, collection)
Query
Virtual Metadata Catalog
Virtual Metadata Attributes and Mappings
Reasoner
Query Mapping
MCS Query
Metadata Catalog
Metadata Attributes
Metadata Catalog
Metadata Catalog Service (MCS)

Figure 1: Architecture of a Virtual Metadata Catalog.

MCS query is constructed by adding the operators to construct the appropriate query formula.

3 Discussion

In prior work we developed Artemis [Tuchinda *et al.*, 2004], a query mediator for metadata catalogs that used semantic representations to integrate several metadata catalogs. Artemis uses a centralized approach with a single reasoner that incorporates all the representations and mappings to all the metadata catalogs. The approach we take in this paper is decentralized in that a reasoner is associated with each metadata catalog. Many projects are developing custom solutions for managing metadata. For example the Collaboratory for Multi-scale Chemical Science (CMCS) has extended the Dublin Core metadata schema to support metadata representing in chemical sciences [Myers *et al* 2004]. Based on this common schema they have developed an interface that allows users within the community to query, browse and register the metadata within a portal environment. The Storage Resource Broker (SRB) [Baru *et al.*, 1998] is a metadata management system that, unlike the work presented here, is based on a centralized metadata catalog (MCAT) and does not provide semantic information about the catalog content. The myGrid project [Wroe *et al.*, 2003] models data sources as semantic web services, and relies on the use of standard ontologies to alleviate the problem of semantic integration.

This work illustrates how semantic representations can be used to support virtual metadata attributes and how reasoners can be used to resolve queries that use them, opening the way for virtual metadata catalog services. As proof of concept, the system implemented so far can answer queries but that is only part of the functionality of a metadata catalog service. We plan to extend its functionality to the fullest in the future and include useful functions such as finding available metadata attributes, publishing data, grouping data into hierarchical collections, and setting authorization information on objects.

Acknowledgements. We gratefully acknowledge the support from the National Science Foundation through the project Towards Cognitive Grids with grant number SCI-0455361 and the SCEC-CME project with grant number EAR-0122464.

References

[Baru *et al.*, 1998] C. Baru, R. Moore, A. Rajasekar, M. Wan. The SDSC Storage Resource Broker. *Proceedings of Proc. CASCON'98 Conference*, Dec. 1998.

[Myers *et al.*, 2004] J. D. Myers et al. “A Collaborative Informatics Infrastructure for Multi-scale Science”, Published in the proceedings of the Challenges of Large Applications in Distributed Environments (CLADE) Workshop, June 7, 2004, Honolulu, HI.

[Singh *et al*., 2003] G. Singh, S. Bharathi**,** A. Chervenak, E. Deelman, C. Kesselman, M. Manohar, S. Patil, and L. Pearlman. A Metadata Catalog Service for Data Intensive Applications, *Proceedings of SC Conference, 2003*.

[Tuchinda *et al*., 2004] R. Tuchinda, S. Thakkar, Y. Gil, and E. Deelman. Artemis: Integrating Scientific Data on the Grid. *Proceedings of the 16th Conference on Innovative Applications of Artificial Intelligence (IAAI),* San Jose, CA, July 2004.

[Wroe *et al.*, 2003] C. Wroe, R. Stevens, C. Goble, A. Roberts, and M. Greenwood. A Suite of DAML+OIL ontologies to describe bioinformatics web services and data. *Journal of Cooperative Information Science*, Vol. 12, No. 2, 2003.