Virtual Metadata Catalogs:
Augmenting Existing 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

*Grid computing provides infrastructure to manage large amounts of data by separating the data itself (and its physical rendering in replicas) from the metadata that describes the nature of the data (often called logical data descriptions). This is particularly important in scientific applications where large, heterogeneous, and distributed collections of data need to be accessed by many users. 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 criteria. However, the query has to be formulated in terms of the metadata attributes defined for the catalog. Our work explores the concept of* virtual metadata, *where catalogs can be queried using metadata attributes not originally defined in the catalog using semantic web standards.*

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 [3,8]. 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 very diverse uses and possibly conceptualizations of the data.

We have developed an approach that augments the existing metadata catalogs with semantic representations to create *virtual metadata catalogs*. Figure 1 illustrates our approach. 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*).

Our implementation of a virtual metadata catalog was developed using MCS [3]. 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 the MCS, which returns all the references to data stored in it that satisfy the query.

In our implementation we used data from three different domains: climate modeling, earthquake science, and workflow execution tracking and performance. Although these various domains deal with different types of data, they all make use of temporal concepts. We used the OWL Time ontology [4] to support the mappings. The Virtual Metadata Attributes and Mappings express that the MCS "startDate" attribute is equivalent to the "from" virtual metadata attribute, and the MCS attribute "endDate" is equivalent to the "to" virtual metadata attribute. These mappings are specified in OWL. The Query Mapping component accepts the OWL Query document and configures the semantic reasoner by loading the OWL ontologies and

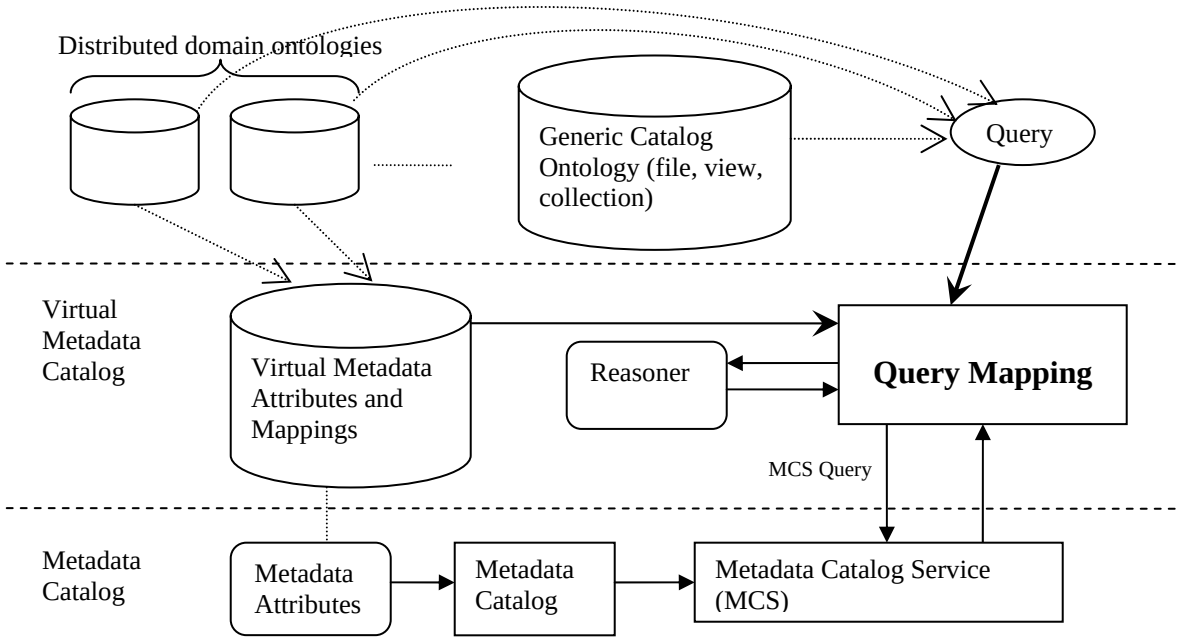

Figure 1: Architecture of a Virtual Metadata

rules referenced in it. Another mapping performed by the system is the conversion of the values from the XML Schema Datatypes to the ones that are expected by the MCS database.

In prior work we developed Artemis [2], 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. The Storage Resource Broker (SRB) [7] is a metadata management system that, unlike the work presented here, is based on a centralized metadata catalog and does not provide semantic information about the catalog content. The myGrid project [10] models data sources as semantic web services, and relies on the use of standard ontologies to alleviate the problem of semantic integration.

In future work, we would like to extend the query mapping process to make it more robust and better integrated with OWL reasoners as well as the MCS back end. We would like to formalize the mappings of different query expressions in a comprehensive framework. This will be facilitated as standard OWL query languages emerge.

References

[2] R. Tuchinda, S. Thakkar, Y. Gil and E. Deelman, “Artemis: Integrating Scientific Data on the Grid”, *IAA-04*, San Jose, CA, July 2004.

[3] 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”, *SC 2003*.

[4] J. R. Hobbs and F. Pan, “An Ontology of Time for the Semantic Web”, *ACM Transactions on Asian Language Processing (TALIP): Special issue on Temporal Information Processing*, Vol. 3, No. 1, March 2004, pp. 66-85.

[7] C. Baru, et al., "The SDSC Storage Resource Broker," Proceedings of Proc. CASCON'98 Conference, 1998.

[9] [10] Wroe, C., R. Stevens, C. Goble, A. Roberts, and M. Greenwood. (2003), "A Suite of DAML+OIL ontologies to describe bioinformatics web services and data". *Journal of Cooperative Information Science.* Vol. 12, No. 2 (2003)