
Enabling Interoperability – Supporting a Diversity of Search Paradigms Using Shared Ontologies and Federated Registries

J. Steven Hughes¹, Dan Crichton¹, Sean Hardman¹, Chris A. Mattmann¹, Paul Ramirez¹

¹Jet Propulsion Laboratory, California Institute of Technology Pasadena, CA 91109, USA

The advent of the Web and languages such as XML have brought an explosion of online science data repositories and the promises of correlated data and interoperable systems. However there have been relatively few real successes since research and experience suggests that just having physical and syntactic connectivity is not adequate. The remaining problem is that the semantic issues, the meaning of the data, must still be addressed. The authors will present an information systems architecture that directly addresses this problem with search applications that are built by harvesting metadata from federated registries where the metadata is compliant to one or more shared ontologies.

Experience suggests that no single search paradigm will meet all search needs in a community. Traditional *forms based* search is still considered critical by a significant percentage of most science communities. However *text base* and *facet based* search are showing great improvements in the overall community perception that search can be easy and that the data is available and can be located. Finally *semantic search* promises ways to find data not conceived when the metadata was first captured and organized. This situation suggests that successful science information systems must be able to deploy new search applications quickly, efficiently, and often for ad-hoc purposes. By harvesting registered metadata that is compliant to shared ontologies the first half of the problem, the metadata management problem, including metadata correlation, is addressed. The issues that remain are software and human interface related.

Federated registries allow data to be packaged or associated with their metadata and managed as simple registry objects. Standard reference models for federated registries now exist that ensure registry objects are assigned globally unique identifiers at registration and object versioning, classification, and cataloging are addressed automatically. Also supported are object security and integrity, subscription and notification of registry events, and federated query, linking, and replication of registry objects across geographically distributed repositories. Finally each registry is locally governed.

Key principles for shared ontology development are that the ontology remains independent of its implementation. The ontology should be allowed to evolve at a different pace than the implementation technology and the ontology should drive both the development and management of the system implementation. The ontology must also be extensible, flexible and scalable. The dichotomy between data and its description and the descriptions of physical and conceptual things in the domain need to be unified under a standard model, such as the Open Archive Information System (OAIS) Information Object. Finally the fact that ontology development is a difficult task that requires time, patience and experts in both the science domain and information modeling must be accepted. Note however that the data driven paradigm allows system development to start as soon as a preliminary shared ontology is available.

The Planetary Data System (PDS) has adopted this architecture for its next generation information system, PDS 2010. The authors will report on progress, briefly describe key concepts, and illustrate how the new system will be phased into operations to handle both legacy and new data. In particular the shared ontology is being used to drive system implementation by generating standards documents and configuration files automatically after each update. The resulting information system including search applications, harvesting tools, federated registries, shared ontologies, and compliant metadata will go a long way toward meeting the expectations of modern scientists by providing the information interconnectedness, correlative science, and system interoperability that they desire, across distributed science data repositories.
