Information Models, Data Requirements, and Agile Data Curation

John S. Hughes (1), Dan Crichton (1), Bernd Ritschel (2), Sean Hardman (1), and Ron Joyner (1)

(1) Jet Propulsion Lab, Pasadena, United States (steve.hughes@jpl.nasa.gov), (2) Helmholtz-Zentrum Potsdam - Deutsches GeoForschungsZentrum - Germany

The Planetary Data System's next generation system, PDS4, is an example of the successful use of an ontology-based Information Model (IM) to drive the development and operations of a data system. In traditional systems engineering, requirements or statements about what is necessary for the system are collected and analyzed for input into the design stage of systems development. With the advent of big data the requirements associated with data have begun to dominate and an ontology-based information model can be used to provide a formalized and rigorous set of data requirements. These requirements address not only the usual issues of data quantity, quality, and disposition but also data representation, integrity, provenance, context, and semantics. In addition the use of these data requirements during system's development has many characteristics of Agile Curation as proposed by Young et al. [Taking Another Look at the Data Management Life Cycle: Deconstruction, Agile, and Community, AGU 2014], namely adaptive planning, evolutionary development, early delivery, continuous improvement, and rapid and flexible response to change. For example customers can be satisfied through early and continuous delivery of system software and services that are configured directly from the information model.

This presentation will describe the PDS4 architecture and its three principle parts: the ontology-based Information Model (IM), the federated registries and repositories, and the REST-based service layer for search, retrieval, and distribution. The development of the IM will be highlighted with special emphasis on knowledge acquisition, the impact of the IM on development and operations, and the use of shared ontologies at multiple governance levels to promote system interoperability and data correlation.