

A Semantic Framework for the Management, Analysis and Assimilation of Mixed-media Scientific Data

Suzanne Little, PhD Thesis, ITEE Dept. University of Queensland, 2006-08-03

Email: little@itee.uq.edu.au

Abstract: Advances in scientific research techniques and technologies have led to an explosion of information-rich, mixed-media data. For example, new high-throughput data capture and analysis techniques involving electron microscopes, tomographic scanners, micro-arrays, satellites and telescopes, has resulted in the generation of research data in quantities that are too great for effective assimilation. This data is produced in a range of mediums and formats, including numerical data, spectrographic output, genomic arrays, images, 3D models, audio and video, in disciplines such as nano-materials, bioinformatics, tele-medicine, geosciences and astronomy. Scientific discovery is driven by the storage, dissemination, analysis and correlation of these complex data sets in the networked digital world of today.

Researchers need tools that support distributed scientific procedures, carried out by collaborative teams and that enable them to securely manage, analyse and assimilate the volume and variety of data generated. The multi-dimensional nature of media exacerbates the difficulty of incorporating newly generated data into existing understandings, models or systems. In addition, scientists need to be able to validate and authenticate scientific results through recording of specific and detailed provenance metadata describing the precise methodology and outcomes. Finally, today's researchers are increasingly working in large collaborative groups spread across multiple organisations, domains and geographical locales. Semantic interoperability is required to overcome these challenges and support distributed querying, analysis and integration of mixed-media, heterogeneous data.

The semantic web promotes interoperability through formal languages and semantics. It aims to build a web where information is exchanged easily between humans and machines. The semantic web architecture is comprised of layered standards and protocols for data definition, storage and exchange (such as eXtensible Markup Language (XML)¹, Resource Description Framework (RDF)², Web Ontology Language (OWL)³, Uniform Resource Identifiers (URIs)⁴). This approach aims to define and expose the semantics associated with data or information, in order to facilitate automatic processing, sharing and reuse of the data.

Accordingly, this thesis aims to apply, evaluate and extend emerging semantic web technologies to provide innovative approaches for semantically annotating, integrating and correlating distributed mixed-media scientific data. To achieve this, I propose a semantic framework that incorporates technologies from the semantic web. This will support interoperability through formal syntaxes, ontologies and inferencing rules. The framework will enable innovative search, data exploration, hypothesis development and evaluation interfaces and will assist researchers in managing, assimilating and distributing data to facilitate further scientific understanding and discovery.

¹ <http://www.w3.org/XML>

² <http://www.w3.org/RDF>

³ <http://www.w3.org/2004/OWL>

⁴ <http://www.w3.org/Addressing/>