

Semantic Data Integration: Answer to Complexity in Translational Research

Erich Gombocz¹⁾, Robert Stanley¹⁾, Jason Eshleman¹⁾, Chuck Rockey¹⁾,

¹⁾ IO Informatics, Inc., Berkeley, CA, USA.

POSTER ABSTRACT

Traditional, relational data warehousing and federation approaches can scale well and are effective for many core data storage and access requirements. However, such approaches often fail when facing the dynamic changes and the inherent complexity of data integration requirements for health care / life science (HCLS) research. The biggest challenge in today's cross-domain research efforts is still the extraction, translation, and loading (ETL) of data from multiple heterogeneous resources for integration into coherent, contextualized relationships. More flexible and extensible solutions are mandated to enable the demands of modern research.

This poster outlines semantic methods which make it possible for domain experts, ontologists and informaticians, to quickly build, modify and extend integrated knowledge bases. These methods fulfill dynamic integration requirements and provide the framework for rich semantic queries (SPARQL) to answer complex biological questions. Through several customer use-cases, the unparalleled power of applying semantic technology to this task is exemplified.

Semantic integration methods assure coherence, harmonize synonyms and different terminologies, and provide an extensible data integration platform and interactive knowledge base for relevant network analysis. This demonstrates the success of using an innovative semantic approach towards integration of all experimental, internal, external, clinical and public data sources. The resulting visual exploration of such an integrated graph environment and the construction of characteristic marker patterns or molecular signatures are applicable to predictive functional biology-based decision support for complex translational research and personalized medicine applications. SPARQL queries can be captured visually and saved in arrays representative for specific biological functions. Being able to create, visualize and test complex models in an easy, automated way makes these methods widely applicable.

The ability to meaningfully integrate and traverse systems-oriented networks rapidly and easily, without losing the underlying complexity, is critical. Establishing concise, actionable inferences about targets, drug interactions, disease states and treatments, using combined clinical, -OMICs and molecular phenotypic data in conjunction with mechanistic insights from public knowledge networks presents a remarkable step forward for translational research. The ability to more efficiently and effectively combine and search data and public knowledge is a stride towards widespread use in patient-centric personalized medicine.