

Working Group on Semantic Applications in Translational Research: Industry and academic centers of excellence join forces to build coherent networks to comprehend complex biological systems in context of personalized medicine and rational drug design

Ted Slater¹⁾, Jonas S. Almeida²⁾, Mark Wilkinson³⁾, Alan J. Higgins⁴⁾, Erich A. Gombocz⁵⁾, Toshiro Nishimura⁵⁾

¹⁾ Pfizer, Inc., Indications and Pathways Center of Emphasis, Chesterfield, MO 63017, U.S.A.

²⁾ Univ. Texas MDAnderson Cancer Center, Houston, TX 77030, U.S.A.

³⁾ The James Hogg ICAPTURE Centre, Vancouver, BC VC6Z1Y6, Canada

⁴⁾ Viamet Pharmaceuticals, Durham, NC 27713, U.S.A.

⁵⁾ IO Informatics, Inc., Berkeley, CA 94710, U.S.A. [Correspondence: egombocz@io-informatics.com]

Summary

This report presents rationales and results of a working group with members from industry and academia in translational research. The Working Group's goals are to develop strategies and toolsets using semantic approaches to better understand multiple data from –OMICS, tissue imaging and clinical trial results as well as public resources within their biological relevancy.

The group, established in July 2007, has been substantial since its inception through its members' contribution towards the mandate utilizing semantic technologies for hypothesis generation in translational research, building a foundation of real-life scenarios to help bridging the gap between observations and understanding of complex relationships in biological systems. The group's successful work on best usage of existing standards efforts while – at the same time – bridging the many differences between semantic formats, taxonomies and ontologies towards a merged knowledgebase is exemplified on examples of integration of experimental datasets with a taxonomy for medical disease codes as well as the development of behind-the-scene application of multiple thesauri. This obtains coherence through controlled vocabularies for both, classes of data and their relationships, and implications are discussed on examples. Pros and Cons of using formal RDF/OWL/OBO/N3-formatted data and output from queries to relational databases meaningfully together are addressed.

We present the resulting prototype methodology, and discuss its application to understanding biological systems using network approaches by means of the examples of multi-biomarker panels for diagnostics in personalized medicine or toxicity, and risk assessment in rational drug design efforts are presented.

Mandate

- Founded in July 2007, the mandate of the working group has been on semantic applications for hypothesis generation in translational research focusing on biomarker identification and qualification as a first tangible step towards a functional semantic model which can account for a better understanding of complex biological network interactions.
- The working group members, experts from industry (Pfizer, Viamet Pharmaceuticals, Cognetics (Division of Clinical Data Inc.)) and academic centers of excellence (MD Anderson Cancer Center; James Hogg ICapture Centre), have shared their expertise and vision with IO Informatics to help developing next generation tools to meaningfully look at coherently integrated data.
- Applying real life use cases, the group was exploring, how one with access to all the data and applications in one place does go forward in exploring and analysing the network and what this can tell researchers and physicians about a disease, target, compound or biomarker. IO Informatics' "Sentient" software provides user-driven and automated methods for data access, analysis, structuring, knowledge building and sharing to create coherent, semantically rich datasets out of previously disconnected or inaccessible data. Objective of the group was to discuss implications and requirements for widely applicable sets of semantic tools across disciplines.
- Through open sharing of ideas, the working group has been substantial in helping to address the challenges outlined below effectively and to develop a methodology for researchers to make sense of all their data in context.

Challenges

- Data coherence: different taxonomies, ontologies, non-standardized vocabularies
- Multiple semantic data formats, several incompatible standards hamper integration efforts
- Ontology merging requires tree-level propagation adjustments for super-classes and sub-classes
- Must be able to map output from relational databases into a relationship-oriented, semantic knowledgebase
- Complexity of network analysis in general: need for easy-to-use tools for scientists
- Scalability of RDF/semantics-based knowledgebases for query and inferencing

Methodology

- Create a textual and graphical data mapper to transform relational database output into semantic data model; save, load and re-apply those mappings for consecutive imports
- Use of one or multiple thesauri for both, classes and relationships upon import warrants data coherence and proper handling of synonyms and relationships according to their meaning
- Source traceability and relationship weighting applied to validate entities; numeric scaling applied to graphical representation
- Account for complexity reduction: at least, at most, exactly n common relationships
- 'Find shortest path' function between two elements in the network; selectability of connection depth from 1 to x-levels connected
- Paging entity browser with sorting by object or relationship
- Graphical, textual and SPARQL query function

Applicability

- Merging and mapping of translational research results from –OMICS, assays, tissue analytics and clinical endpoints into a semantic framework to visualize, explore and analyze data in context to their biological function and underlying mechanism
- Use of gene expression profiling, genotyping of patients and clinical test profiles to optimize therapeutic effect and minimize toxicity in disease treatment
- Perform graphical and SPARQL queries on biomarker panels, then re-plot the resulting sub-networks for qualification

Examples and Use Scenarios

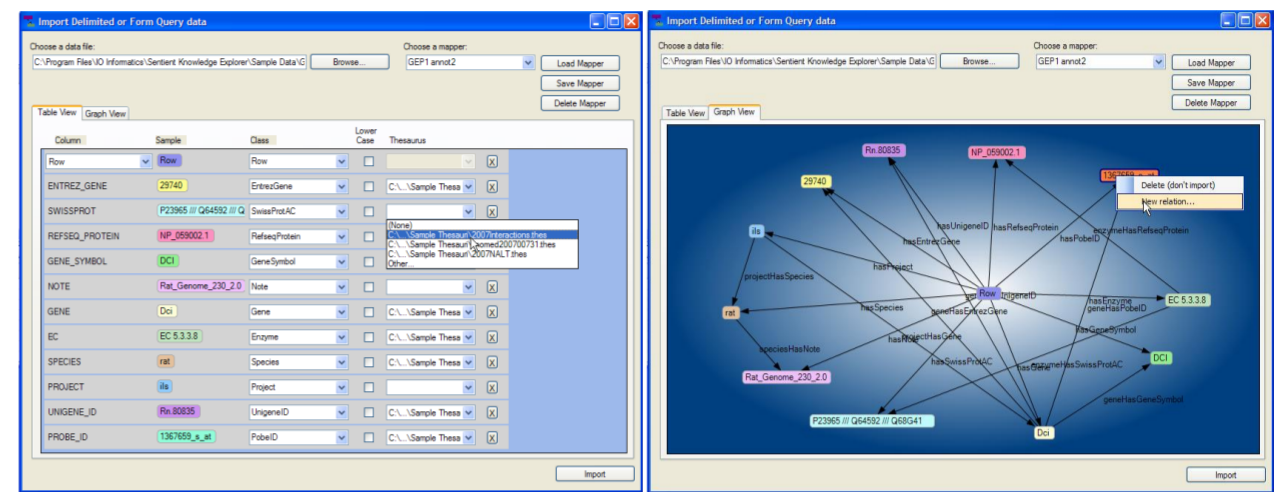


Fig.1: Relationship mapping in Sentient Knowledge Explorer™:

Left: Standardized class definitions using thesauri. Mappers can be applied automatically in the background
Right: Graphical relationship editor allows for authoring to add or remove relationships

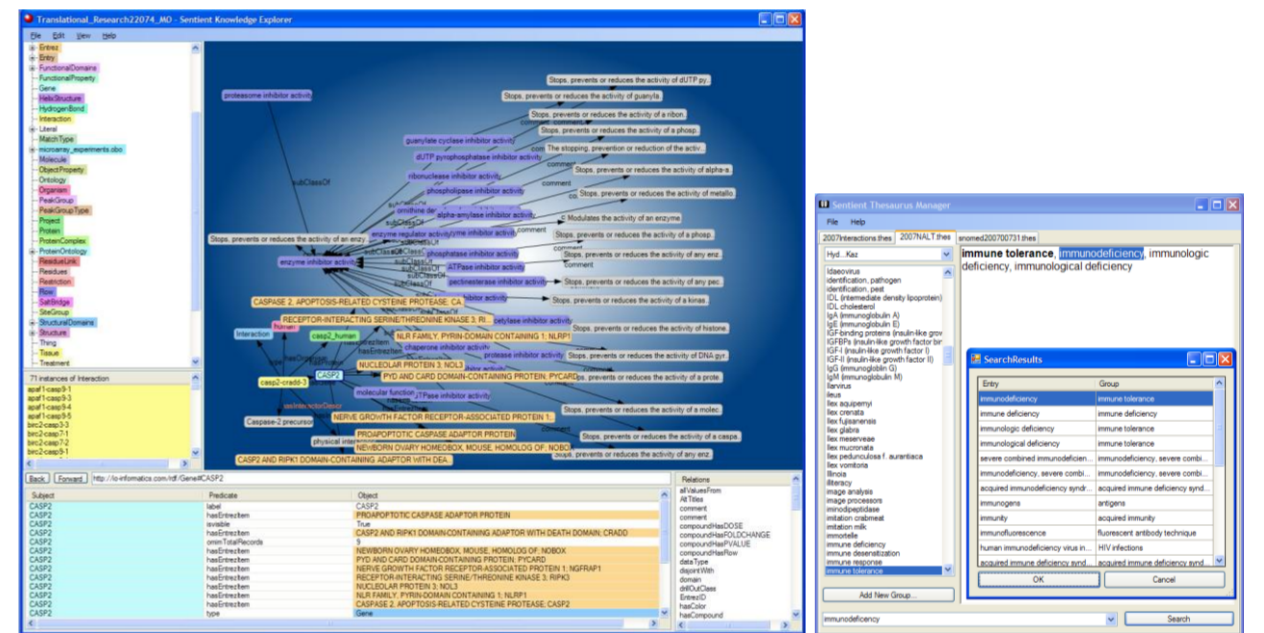


Fig.2: Addressing multiple standards and conflicting data models:

Left: A common knowledgebase from GO-RDF, OWL, OBO, N3, RDF-experimental data and NCBI Entrez resources
Right: Authoring & search in Thesaurus Manager provides for adaptation of representative terms and relationships

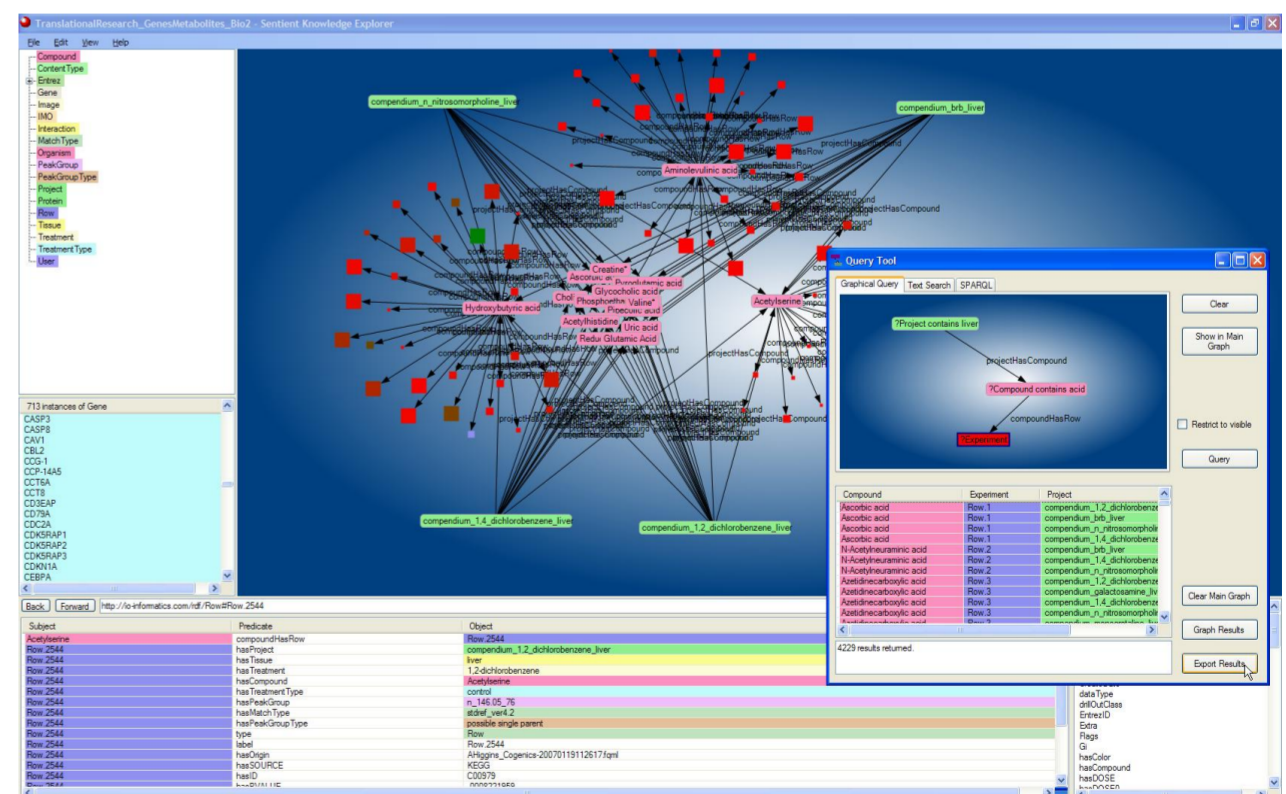


Fig.3: Use case biomarker qualification for rational drug design and personalized medicine:

Toxic responses for different hepato-toxicants: a handful of commonly perturbed metabolites (pink) are potential toxicity biomarkers; the colors (red to green) indicate fold-changes, the square area doses; all timepoints are shown.

Conclusions

Through its expertise and open exchange of ideas, the Working Group was able to

- Provide insights on requirements and suggestions for implementation towards a semantic knowledgebase which now can accept data from multiple different standards (RDF, OWL, OBO, N3) via files, from databases or directly via URIs
- Improving ontology merging and data mapping in the Sentient Knowledge Explorer using one or multiple thesauri for controlled vocabularies for classes and relationships
- Creating a tool for researchers which through its complexity reduction capabilities accounts for functional network analysis to better understand underlying biological processes.