Semantic Data Integration: Answer to Complexity in Translational Research

Erich Gombocz[1), Robert Stanley[1)], Jason Eshleman[1)], Chuck Rockey[1]

[1] IO Informatics, 2550 Ninth Street, Suite 114, Berkeley, CA 94710, USA [Correspondence: egombocz@io-informatics.com]

Summary

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 integration of data from multiple heterogeneous resources into coherent, contextualized information with their 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.

Challenges

- Data coherence: sources, taxonomies, ontologies, non-standardized vocabularies
- Complexity in meaningful integration, scale and dynamics of multiple resources
- Lack of intuitive, science-driven tools for ontology building and hypothesis generation
- Demanding classifier qualification with mechanistic and functional biology insights
- Unifying public resources and internal datasets with proper weighing of knowledge for system models is non-trivial

Approach & Methodology

- Integrate multiple heterogeneous sources (files, spreadsheets, instrument outputs, clinical observations) via intuitive mapping tools into a standards-based semantic framework with contextualized relationships.
- Utilize thesauri to harmonize classes, relationships and data instances during import, preserve coherence through preferred terms and maintain provenance information.
- Create application ontologies dynamically, and/or merge it with (entire or parts of) formal ontologies from public resources such as NCBO or OBO.
- Enrich experimental correlation networks through incorporation of public reference and mechanistic resources, either by direct import or via queries to SPARQL endpoints.
- Explore the resulting network graph for intersections and/or exclusions of relevant characteristics to construct marker pattern for specific biological functions; qualify and refine the model iteratively.
- Save arrays of validated SPARQL query pattern (molecular signatures) in an "Applied Semantic Knowledgebase (ASK)".
- Apply patterns for decision-supported screening with scoring of "hit-to-fit" between model and patient response via secure web portal or Smart Phone.

Results

- Translational research in HCLS has unusually dynamic needs for data integration, applications and search workflows. Through their far more flexible methodologies, semantic technologies are uniquely suited to fill these needs.
- When data are represented as system network, crucial contextual relationships between all information are maintained; queries can take advantage of inference and reasoning.
- 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.
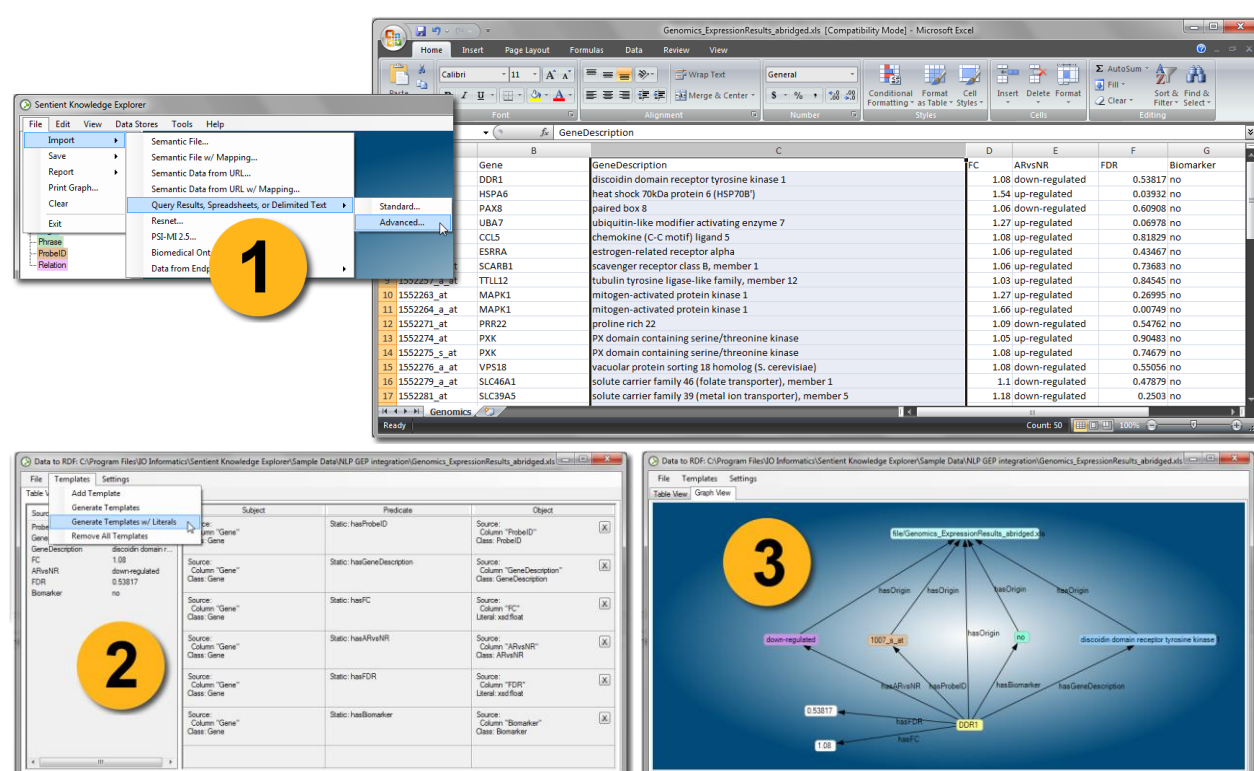

Fig. 1: **Semantic data integration at work**
Intuitive advanced mapping tool for standards-compliant RDF generation from delimited text, spreadsheets or relational database output. Select spreadsheet (Step 1), create mapping (Step 2) and review it (Step 3).
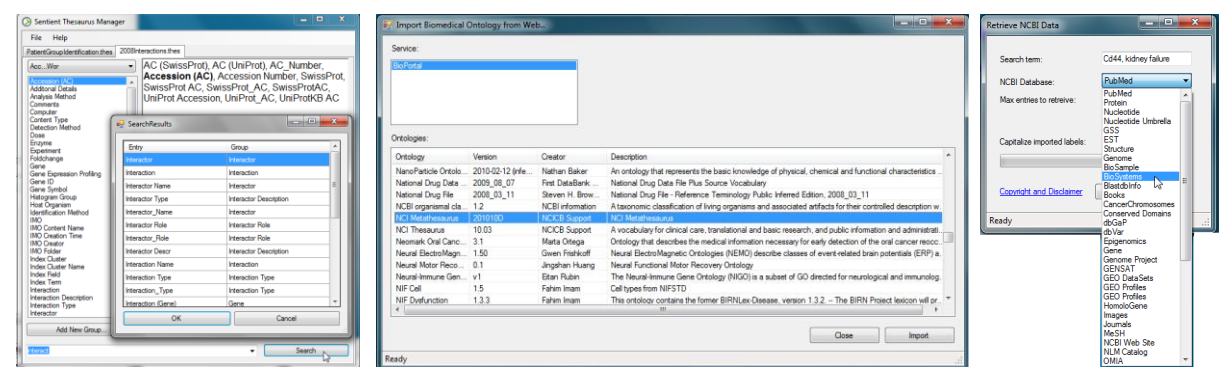

Fig. 2: **Coherent import and data merging across domains**
Harmonization with multiple thesauri (left), ontologies from NCBO's BioPortal (center) and data from Entréz/NCBI (left) are directly applicable for import and merge at integration stage.
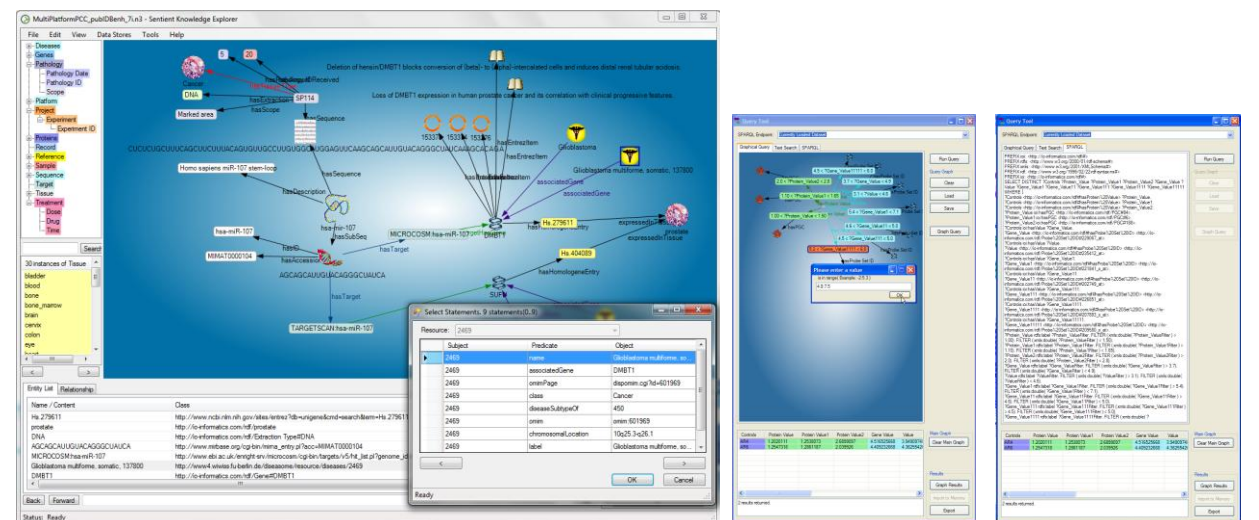

Fig. 3: **Enrichment of experimental correlation networks with public knowledge networks**
Mechanistic qualification of functional biology aspects: Assert classifier relevancy using import from SPARQL endpoints, via web service connections or through graphical query.
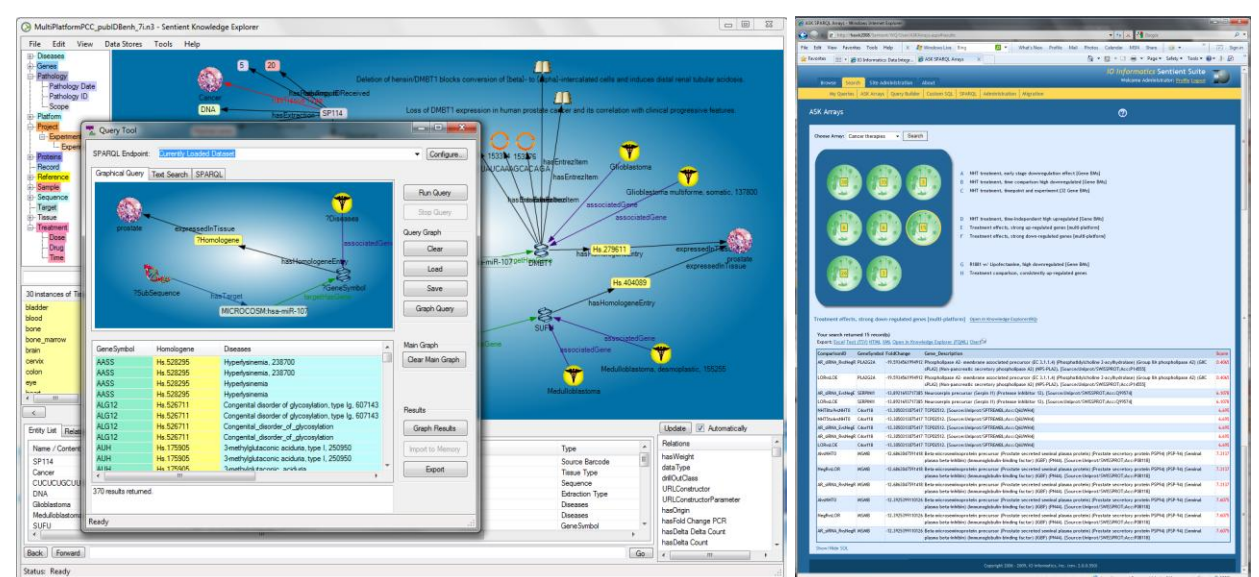

Fig. 4: **Use case of an integrated systems biology network for predictive biology**
Representative systems biology network with molecular signature represented as visual SPARQL query (left), web-based screening for effectiveness of combinatorial cancer therapies based on patient profiles (right).

Discussion

- Translational research requires flexible, dynamic integration in rapidly evolving data environments. Data integration in cross-domain research efforts into coherent, contextualized relationships is still an unmet need in the life sciences.
- Semantic technology-based network models are flexible, more efficient and faster than conventional relational methods, which are well suited for scale, but not for complexity.
- Knowledge built from all available internal and public data helps gaining insights into intricate biological mechanisms and provides a systems-based decision support environment for pharmaceutical industry, life sciences researchers and clinicians alike.
- Actionable inferences about diseases, treatments or other complex biological questions using combined multi-OMIC and molecular phenotypical data enriched with mechanistic insights from public knowledge networks represent a remarkable step towards the understanding needed in knowledge-based patient-centric translational medicine applications.

Implications

- Using all available resources in context, semantic data integration provides effective, dynamic systems-based network model-derived biological signatures to answer complex translational research science questions.
- Addressing complexity via Applied Semantic Knowledgebases (ASK™) provides researchers in life sciences and clinics a novel way to confidently answer burning research questions today.

Outlook & Conclusions

- Concise information which data (targets, compounds, diseases, and patients) fit the model, and how good the match is in each particular case is changing the way knowledge is built, refined and applied.
- As such, it builds the foundation for next generation research and tomorrows healthcare systems.

References

[1]) K. Qaadri:"Knowledge building environments", Best Practices for Personalized Medicine (B2PM 2011), Vancouver, BC, Canada (2011).

[2]) E. Gombocz, D. Milward, J. Eshleman: "Contextual understanding of experimental data via formal semantic integration of NLP-extracted content with other semantically integrated resources", Conference on Semantics in Healthcare and Life Sciences (CSHALS 2011) , Cambridge, MA, USA (2011).

[3]) C. Rockey: "A 'Killer App' for Semantic Technologies: Point-and-Click Data Integration Tools Make it Easy to Deliver Targeted Semantic Knowledge Bases", Conference on Semantics in Healthcare and Life Sciences (CSHALS 2011), Cambridge, MA, USA (2011).

[4]) R. Stanley, E. Gombocz, J. Eshleman, C. Rockey: "From Concepts to Production: Semantic Technology Solves Real Life Sciences and Healthcare Challenges", Semantic Technology 2010 (SemTech), San Francisco, CA, USA (2010).

Acknowledgements