

# From Concepts to Production: Semantic Technology Solves Real Life Sciences and Healthcare Challenges

Robert Stanley<sup>1)</sup>, Erich Gombocz<sup>1)</sup>, Jason Eshleman<sup>1)</sup>, Chuck Rockey<sup>1)</sup>

<sup>1)</sup>IO Informatics, Inc., 2550 Ninth Street, Suite 114, Berkeley, CA 94710, U.S.A. [Correspondence: [rstanley@io-informatics.com](mailto:rstanley@io-informatics.com)]

## Summary

While semantic technologies have been around for quite some time, real proof of being able to deliver on its promise has been lacking and therefore life science companies and clinical centers of excellence have been reluctant to implement it. With the advent of translational research and science spanning multiple information domains, a dynamic, flexible, and extensible solution has become more necessary than ever to cope with the demand for knowledge management and data sharing across disciplines. As data from countless diverse sources (new instruments, files, images, database, and web content) are generated in ever-changing formats, new approaches are needed. This poster outlines how the transition to a semantic data integration, where data are in context with other data, has changed the playing field.

Using real-life challenging scenarios (a. Discovery of genomic & proteomic biomarkers for toxicity classification in a NIST/CLDA setting, b. Development of species-independent disease markers for FDA VET to minimize animal experiments, c. Compound purity and excipient influence on drug stability at Pfizer, d. Comparative treatment effectiveness assessment for prostate cancer at the Prostate Cancer Centre at UBC, and e. Predictive decision support for organ failure in transplant patients at the PROOF Centre for the Prevention of Organ Failures) we will demonstrate, that semantic technology not only provides the toolset required for the new information age in life sciences and clinics as it harmonizes synonyms and nomenclature, but also allows for relationship mining, inference, and reasoning in a systems approach – leading to informed decision-making with high confidence.

The presented solutions using dynamic, flexible and scalable RDF triples stores, SPARQL endpoints, and "linked open data" initiatives are key in research demanding global collaborations. Graph representations of merged data correlation networks and mechanistic reference networks allow identifying classifiers in biomarker discovery and focusing on relevant dimension-reduced subnetworks. Knowledge from powerful systems biology-based models fosters understanding of complex biological processes (e.g., diseases & disease states, pre-disposition to certain biological responses, patient stratification for trials and treatment, predictive risk assessment for tumor growth, organ rejection, and severity of drug side effects).

These successfully demonstrated implementations are testimonials for capabilities, power and effectiveness of semantic technologies to cope with "imprecise connections" across multiple sources effectively, in less time, and at lower costs as compared to alternative approaches. Real peoples' real problems have been solved with real technology.

## Challenges

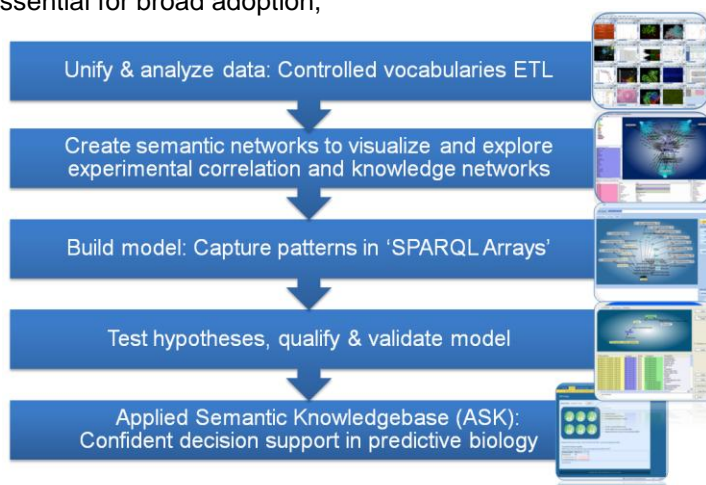
- Complexity of coherent data integration: different types, sources, taxonomies, ontologies, and non-standardized vocabularies.
- In many cases, data relationships are not *a priori* contained in the data sets
- Experimental correlation networks do not necessarily align with mechanistically driven functional biology.
- Reluctance in implementation of relatively new semantic tools for network graph analysis and query in life sciences and clinical production environments.

## Considerations

- Modern triples stores resolve most previous scalability and performance issues.
- Security, provenance, regulatory compliance (HIPAA) need to be addressed.
- Ease of use of tools is essential for broad adoption,

## Methodology

- Map to a common local / formal ontology
- Explore network, reduce its complexity
- Build model using arrays of SPARQL
- Test and refine model
- Use Applied Semantic Knowledgebase for screening



## Results

- **Toxicity classification:** Genomic and metabolic markers to detect different types of toxicity and SPARQL arrays characterizing them have been developed.
- **Species-independent disease markers:** Genomic, proteomic and imaging endpoints have been analyzed across different animal species for biomarker discovery of species-independent disease markers applicable to human diseases
- **Effective treatment with minimal side effects:** Multi-platform genomic and proteomic approaches have been undertaken to characterize the effectiveness of prostate cancer treatment.
- **Purer, more stable drug formulations:** Semantic integration of multiple data sources across imprecise connections allows assessing impact of excipient and compounding purity on stability of drug formulations.
- **Life-threatening organ failure prevention:** Screening of heart transplant patients for likelihood of organ failure bases on combinatorial biomarkers.

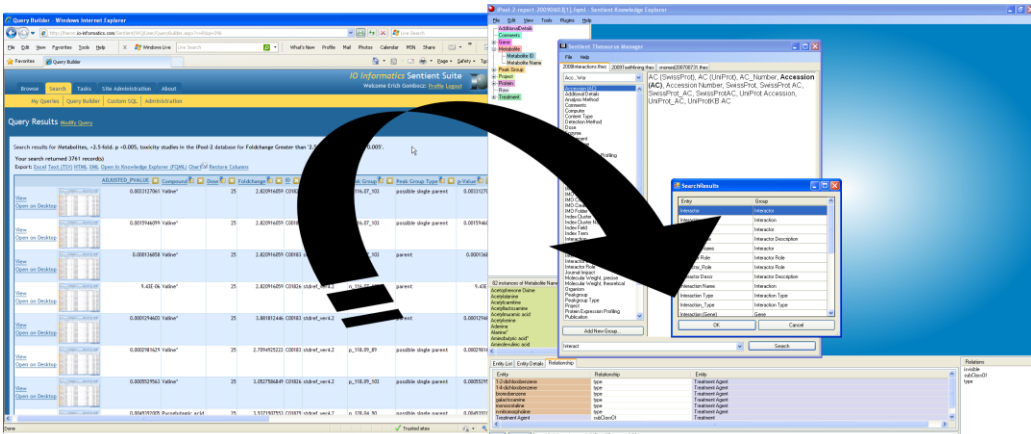


Fig. 1: Data integration framework: From multiple data sources to an experimental network

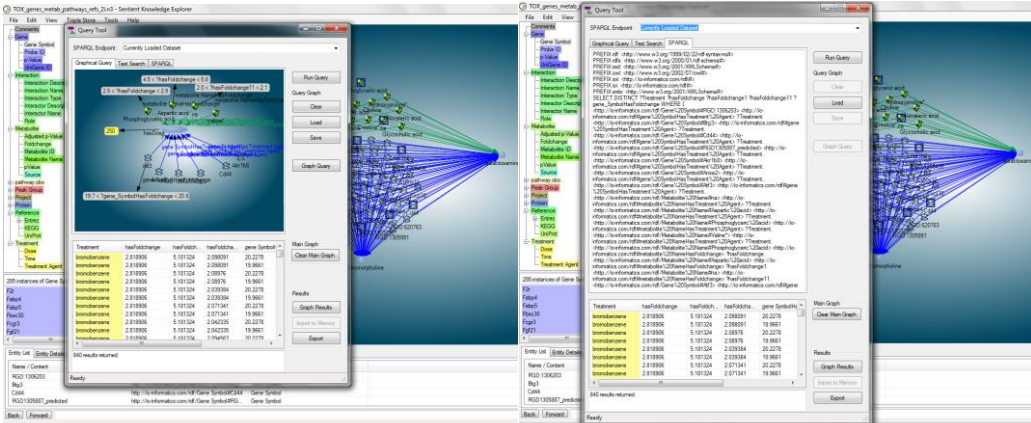


Fig. 2: Toxicity biomarker model refinement (Visual SPARQL)

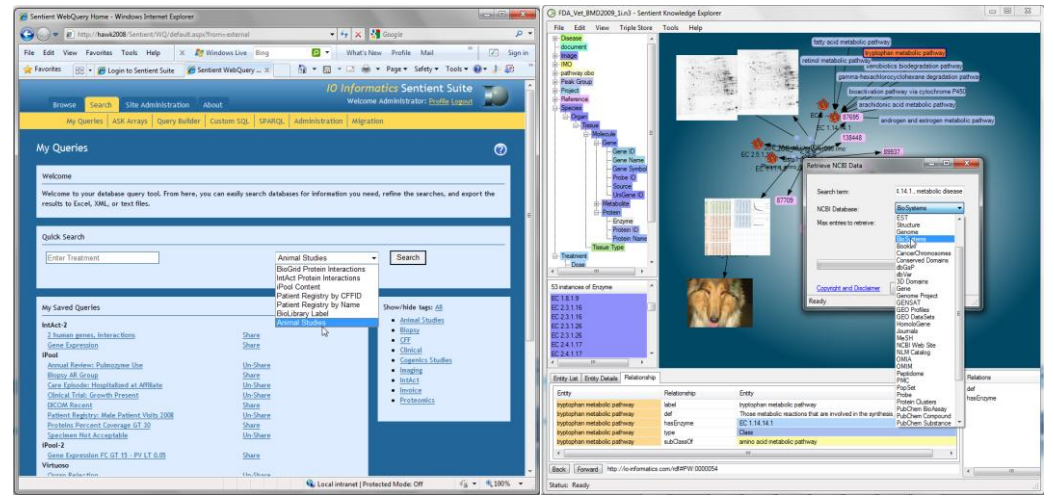


Fig. 3: Species-independent disease markers

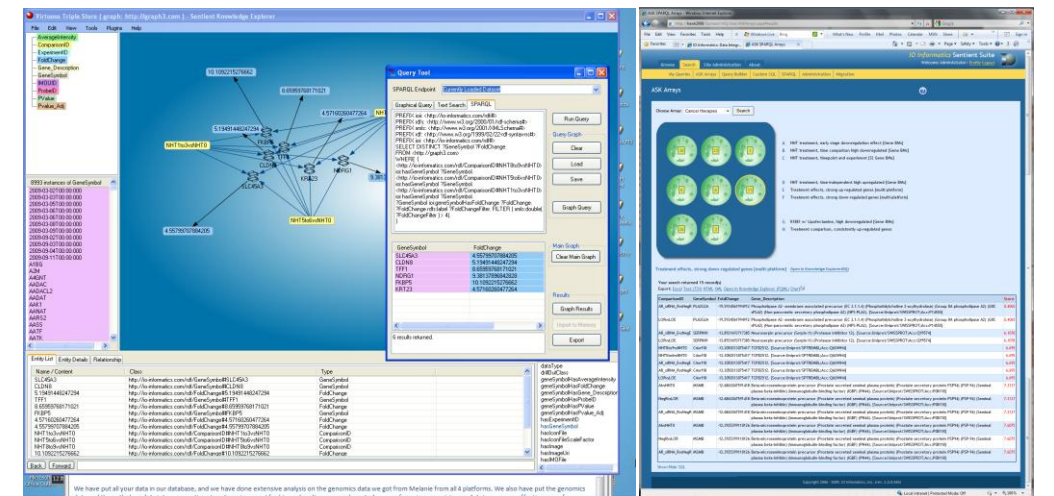


Fig. 4 Comparative prostate cancer treatment effectiveness

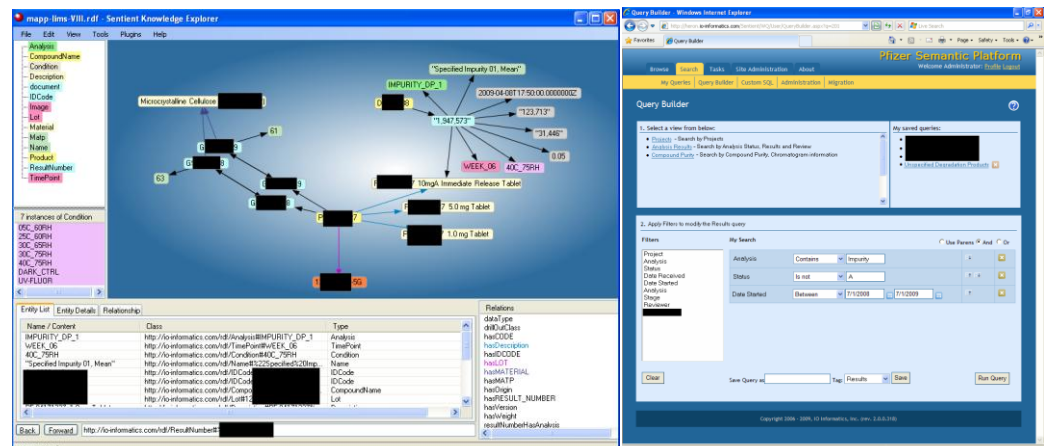


Fig. 5: Compound purity and excipient influence on drug stability



Fig. 6: Decision support for likelihood of organ failure in transplant patients

## Discussion

Applying semantic technology to integration of data and knowledge networks results in:

- Faster, less expensive, dynamic and extensible solutions.
- Meaning, reasoning and inference provide actionable knowledge.
- Interconnected data make research more effective: information access keeps pace with data generation.
- "The Big Picture" from all data in context accounts for more accurate predictive models for complex systems-based functions.
- "Query by Meaning" changes the way we search: imprecise connections can be used to infer non-obvious relationships.
- Linked data collaborations drive knowledge expansion in life & medical sciences.

## Implications for Life Sciences / Healthcare IT

The presented cases exemplify the immediate value across the full cycle from bench to clinic and back to pharmaceutical development. Early failing in drug discovery (biomarkers for toxicity classification), reduction of animal experiments and confident applicability to characterize human diseases (species-independent disease markers), effective treatment with minimal side effects (comparative cancer therapies), purer, more stable drug formulations (excipient influence on impurities), and life-threatening organ failure prevention (transplant rejection) commonly save time, money and lives. They establish widely applicable sustainable and future-proof solutions. More generally,

- The understanding of biological mechanisms provides effective, patient-centric personalized medicine.
- The technology is broadly applicable in life sciences and healthcare.
- Semantic technology-based systems are now production solutions, not just research.

## Acknowledgements

Broad toxicity studies were conducted under NIST Advanced Technology Program (ATP), Award # 70NANB2H3009 as a Joint Venture between Icoria / Cogenics (Division of CLDA) and IO Informatics. Microarray studies were conducted under NIEHS contract # N01-ES-65406. This work was also made possible through many contributions from members of IO Informatics' Working Groups on "Semantic Applications for Translational Research" and "Informatics for Personalized Medicine".

## References

- 1) E. A. Gombocz, A. J. Higgins, P. Hurban, E. K. Lobenhofer, F. T. Crews, R. A. Stanley, C. Rockey, T. Nishimura: "Does network analysis of integrated data help understanding how alcohol affects biological functions?" - Results of a semantic approach to biomarker discovery (2008)
- 2) E. A. Gombocz, Z. Rhoades: "Predictive Toxicology: Applied Semantics with major implications towards safer drugs" (2009)
- 3) Z. Rhoades, E. A. Gombocz: "Charting the Unknown: Capturing and Delivering Value From Understanding Complex Biological Responses" (2010)
- 4) E. Gombocz, R. Stanley, J. Eshleman, Z. Rhoades: "From multiple biomarkers to patient-centric personalized medicine: An 'Applied Semantic Knowledgebase' for decision support" (2010)