CSHALS Conference on Semantics in Healthcare and Life Sciences
February 27 – March 1, 2013
Royal Sonesta Hotel Boston, Cambridge, MA, USA.

Yes, we can! Lessons from Using Linked Open Data (LOD) and Public Ontologies to Contextualize and Enrich Experimental Data

Erich A. Gombocz[1], Andrea Splendiani[2], Mark A. Musen[3], Robert A. Stanley[1], Jason A. Eshleman[1]

[1]IO Informatics, Inc., Berkeley, CA, USA
[2]IO Informatics, Inc., London, UK
[3)]Stanford Center for Biomedical Informatics Research (BMIR), Stanford, CA, USA

Correspondence: egombocz@io-informatics.com

Topic Area: **Linked Data**

SUMMARY

Semantic W3C standards provide a framework for the creation of knowledge bases that are extensible, coherent, interoperable, and on which interactive analytics systems can be developed. An ever growing number of knowledge bases are being built on these standards— in particular as Linked Open Data (LOD) resources. The availability of LOD resources has received increasing attention and use in industry and academia.

Using LOD resources to provide value to industry is challenging, however, and early expectations have not always been met: issues often arise from the alignment of public and experimental corporate standards, from inconsistent namespace policies, and from the use of internal, non-formal application ontologies. Often the reliability of resources is problematic, from service levels of LOD resources and/or SPARQL endpoints to URI persistence. Furthermore, more and more "Open data" are closed for commercial use, and there are serious funding concerns related to government grant-backed resources.

With these challenges, can Semantic Web technologies provide value to Industry today?
We make the case that, *yes, this can be done* and is the case now.

We demonstrate a use case of successful contextualization and enrichment of internal experimental datasets with public resources, thanks to outstanding examples of LOD such as UniProt, Drugbank, Diseasome, SIDER, Reactome, and ChEMBL, as well as ontology collections and annotation services from NCBO's BioPortal.

We show how, starting with semantically integrated experimental results from multi-year toxicology studies performed on different platforms (gene expression and metabolic profiling), a knowledge base can be built that integrates and harmonizes such information, and enriches it with public data from UniProt, Drugbank, Diseasome, SIDER, Reactome, and NCBI Biosystems. The resulting knowledge base facilitates toxicity assessment in drug development at the pre-clinical trial stage. It also provides models for classification of toxicity types (hepatotoxicity, nephrotoxiciy, toxicity based on drug residues) and offers better a priori determination of adverse effects of drug combinations. In this specific use case, we were not only able to correlate responses across unrelated studies with different experimental models, but also to validate system changes associated with known common toxicity mechanisms such as oxidative stress (Glutathione metabolism), liver function (Bile acid and Urea cycle) and Ketoacidosis. Since experimental observations from multi-modal –OMICs data can result from the same perturbation, but represent very different biological processes, and because pharmacodynamic correlations are not necessarily functionally linked within the biological network and genetic and metabolic changes may occur at lower doses and prior to pathological changes, enrichment with LOD resources offers new insights into mechanisms and led to discovery of new pharmacodynamically and biologically linked pathway dependencies.

As LOD resources mature, more reliable information is becoming publicly available that can enrich experimental data with computable descriptions of biological systems in ways never anticipated before and that ultimately help in understanding the experiments' results. The time and money saved from such an approach has enormous socio-economic benefits for drug companies and healthcare alike.

As a community, we need to establish business models through cooperation between industry and academic institutions that support the maintenance and extension of invaluable public LOD resources. Their effective use in enriching toxicology data exemplifies the success of using Semantic Web technologies to contextualize experimental, internal, external, clinical and public data towards faster and better understanding of biological systems and, as such, more effective outcomes in health and quality of life for all of us.

Poster structure:

- Summary
- Materials & Methods
- Results
- Conclusions
- Acknowledgements
- References

2 Tables, 3 Figures.

Acknowledgements

Hepatotoxicity studies were carried out under NIST Advanced Technology Program Award#**70NANB2H3009** in a multi-year joint project between Icoria / Cogenics and IO Informatics.
Alcohol studies were performed under NIAAA contract #**HHSN281200510008C** in collaboration with Bowles Center for Alcohol Studies (CAS) at UNC.
The W3C HCLS SIG is acknowledged for many helpful discussions on LOD resources, public SPARQL endpoints and Bio2RDF developments, and Trish Whetzel at NCBO for her support on ontology services integration

References

(1) LDOW2012 Linked Data on the Web. Bizer C,Heath T, Berners-Lee T, Hausenblas M. WWW Workshop on Linked Data on the Web, 2012 Apr.16, Lyon, France.
(2) The National Center for Biomedical Ontology. Musen MA, Noy NF, Shah NH, Whetzel PL, Chute CG, Story MA, Smith B. J Am Med Inform Assoc. 2012 Mar-Apr; 19 (2): 190-5
(3) BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, Musen MA. Nucleic Acids Res. 2011; 39 (Web Server issue): W541-5
(4) Using SPARQL to Query BioPortal Ontologies and Metadata Salvadores M, Horridge M, Alexander PR, Fergerson RW, Musen MA, and Noy NF. International Semantic Web Conference. Boston US. LNCS 7650, pp. 180195, 2012.
(5) The Translational Medicine Ontology and Knowledge Base: driving personalized medicine by bridging the gap between bench and bedside. Luciano JS, Andersson B, Batchelor C, Bodenreider O, Clark T, Denney CK, Domarew C, Gambet T, Harland L, Jentzsch A, Kashyap V, Kos P, Kozlovsky J, Lebo T, Marshall SM, McCusker JP, McGuinness DL, Ogbuji C, Pichler E, Powers RL, Prud'hommeaux E, Samwald M, Schriml L, Tonellato PJ, Whetzel PL, Zhao J, Stephens S, Dumontier M. J.Biomed.Semantics 2011; 2(Suppl 2):S1
(6) VoID Vocabulary of Interlinked Datasets. Cyganiak R, Zhao J, Alexander K, Hausenblas M. DERI, W3C note 6-Mar-2011
(7) PROV-O: The PROV Ontology. W3C Candidate Recommendation 11- Dec-2012
(8) PAV 2.0 – Provenance Authoring and Versioning ontology Ciccarese P. Dec-2010.
(9) From Individual Experiments to Informed Decision Making: Challenges, Sucess Stories and Opportunities in Collaborative Science. Gombocz EA in: Data for Decision Making: Lab, Enterprise, Web, Center for Computing for Life Sciences SFSU. 2012 May 3, San Francisco, CA.
(10) Does network analysis of integrated data help understanding how alcohol affects biological functions? - Results of a semantic approach to biomarker discovery. Gombocz EA, A.J. Higgins AJ, Hurban P, Lobenhofer EK, Crews FT, Stanley RA, Rockey C, Nishimura T. 2008 Sept.29-Oct.1.Biomarker Discovery Summit, Philadelphia, PA.