SEMANTIC TECHNOLOGY & BUSINESS CONFERENCE
June 2-5, 2013 · San Francisco Presented by SemanticWeb.com and WebMediaBrands

"Semantics-in-a-box"

Integrated Data Appliances to contextualize experiments with a world of public knowledge

Erich Gombocz
IO Informatics, Inc., Berkeley, CA, USA

Abstract

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. A growing number of knowledge bases are being built on these standards— in particular as Linked Open Data (LOD) resources, and their availability has received increasing attention in industry and academia. Using LOD resources to provide value to industry is challenging, however, and early expectations have not always been met: issues arise from the alignment of public and experimental corporate standards, from inconsistent URI policies, and from the use of internal, non-formal application ontologies. To add to this, often the reliability of resources is problematic, from service levels to SPARQL endpoint uptime to URI persistence. Not the least, in many cases provenance issues have not properly resolved, and there are serious funding concerns related to government grant-backed resources.

For this reasons, an integrated data appliance (iDA) preloaded with semantically integrated public knowledgebases provides an enterprise-ready "Semantics In-a-box" solution to address those shortcomings effectively. As public datasets exist in many revisions over time, registered and mirrored on many places, with registries often out of date or containing conflicting information, several initiatives have been currently proposed at the W3C and in consortia and industry alliances to align interlinked datasets (such as using vocabulary of interlinked datasets, VoID or PROV-O). For the end user, the dilemma of having to deal with such obstacles as additional non-trivial data mapping as well as the need to have rich authoring, licensing, provenance and versioning (such as developed in PAV) included with the data creates another barrier in broad application of semantically contextualized, integrated experimental and public datasets.

This can be remedied. Using an iDA on a preconfigured enterprise-ready hardware containing semantically integrated sets of public knowledgebases out-of-the-box and providing controlled versioning and maintenance cycles solves this predicament. Integrated client and web applications to visualize explore and query the RDF graphs from a common UI reduce barriers to entry for end users and focus primarily on its scientific utility.

By means of such an approach to better understanding and characterization of toxicity, we show how, starting from semantically integrated experimental results from multi-year toxicology studies performed on different platforms (genomic and metabolic profiling), iDA-hosted public life sciences resources (UniProt, Drugbank, Diseasome, SIDER, Reactome, NCBI Biosystems) can be used to provide models for classification of toxicity types in pre-clinical settings. Due to already pre-aligned RDF with detailed and accurate provenance and versioning, a better a priori determination of adverse effects of drug combinations can be achieved much faster and at much less effort. Rich SPARQL queries allowed to quickly correlate responses across unrelated studies with different experimental models, and to validate system changes associated with known common toxicity mechanisms.

The time and money saved from such an approach has huge socio-economic benefits for drug companies and healthcare alike. Having linked data available in one appliance together with experimental

results makes it easy to employ Semantic Web technologies worry free, and, as such, to promote a better understanding of biological systems more readily..

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) VoID Vocabulary of Interlinked Datasets. Cyganiak R, Zhao J, Alexander K, Hausenblas M. 6-Mar-2011
(5) PROV-O: The PROV Ontology. W3C Candidate Recommendation 11- Dec-2012
(6) PAV 2.0 – Provenance Authoring and Versioning ontology Ciccarese P. Dec-2010.

About the Speaker

Dr. Erich Gombocz has over 30 years experience in Life Science research, laboratory automation and data management in distributed environments, plus more than 30 years programming experience in instrument control, user interface, database design, scientific analysis, on-line laboratory automation, and innovative software architecture.

From 2000-2003, as Chief Science Officer of Biosentients, Erich contributed substantially to design and implementation of the company's informatics technology, including its systems architecture and analytical modules. Focusing on semantic data integration and knowledge management in life sciences, he founded IO Informatics in 2003 together with Bob Stanley to apply systems biology approaches to contextualize knowledge for pharmaceutical and clinical decision-making.

Dr. Gombocz has published over 60 scientific publications and holds currently more than 40 biotechnology- and software-related US and international patents. He is an international expert in separation science and bioinformatics, a member of several professional organizations, and serves on the editorial board of a number of scientific journals. His activities in World Wide Web Consortium (W3C) HCLS, at the National Center for Biomedical Ontology (NCBO) and the Pistoia Alliance Standards Initiative and as Chair of Working Group for Best Practices in Data Sharing are a testimonial for his role at the forefront of technology.