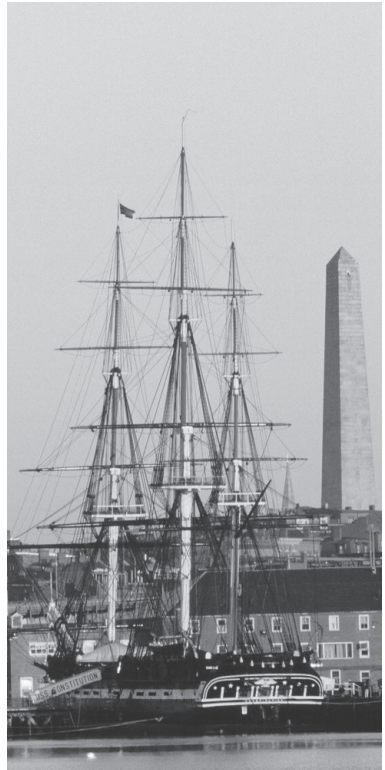




Conference on
Semantics in
Healthcare
And
Life
Sciences

February 23–25, 2011
ROYAL SONESTA HOTEL BOSTON
CAMBRIDGE/BOSTON, USA





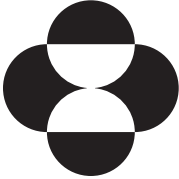
Contents

Sponsors	2
Welcome	3
Schedule	5
Keynotes	8
Tutorial	14
Presentations	16
TECH TALK 1	23
TECK TALK 2	24
TECH TALK 3	37
TECH TALK 4	38
Posters	44



SPONSORS

GOLD



MERCK

NON-PROFIT



Rensselaer

GENERAL



O'REILLY®



An official conference of the
International Society for
Computational Biology



Dear CSHALS participant,

Welcome to the Fourth Annual ISCB Conference on Semantics in Healthcare and Life Sciences!

This conference focuses on the specific applications of semantic technologies within the pharmaceutical and healthcare industries. Your presence indicates you are involved or deeply interested in this cutting edge technology. The conference format is designed to encourage you to engage our visionary speakers, and one another, to improve your understanding and take home new ideas, new methods and new contacts to draw from in the future.

For those among you who might be new to using intelligent information technologies in pharmaceutical R&D, I hope you took advantage of the pre-conference tutorials that were organized by W3C's Semantic Web for Health Care and Life Sciences Interest Group (HCLSIG), and by the Rensselaer Polytechnic Institute's Tetherless World Constellation (RPI). ISCB and W3C have partnered on the tutorial part of this conference each year, and we were pleased to welcome RPI's hands-on approach as a highly valuable element of the overall CSHALS event.

This year's conference was once again organized under the leadership of conference chair Ted Slater, together with organizing committee members Jonas Almeida, Mike Bevil, Lee Feigenbaum, Joanne Luciano and Eric Neumann. ISCB's Director of Conferences, Steven Leard, worked out all logistical arrangements to translate the organizing committee's vision into a cohesive program that fosters discussion and information sharing among all attendees. And, BJ Morrison McKay, ISCB Executive Officer, worked closely with Steven and the organizing committee to support their efforts and ensure that ISCB continues to provide high quality meetings to our members and scientific community. I believe each of these individuals have helped advance ISCB's mission, and I thank them for their commitment to the success of this conference.

Ultimately, the CSHALS experience is yours to shape by contributing to discussions and taking full advantage of the opportunities to network. I thank each and every one of you for attending, and hope that we meet your needs and exceed your expectations.

Enjoy!

Burkhard Rost
ISCB President



Welcome!

On behalf of the entire CSHALS 2011 Organizing Committee, I would like to thank all the speakers for agreeing to present at the Fourth Annual Conference on Semantics in Healthcare and Life Sciences. This conference will be a unique forum for the presentation and discussion of key topics in the rapidly-growing area of semantic information technologies and their practical applications to life sciences and pharmaceutical R&D. We have structured this conference to be an open and exciting experience for all attendees, and your thought-provoking presentations will make this possible. Our intended outcome is to engage all attendees to actively participate, and to identify where technologies are proving successful and where they still need to be developed to meet changing and challenging scientific and business objectives.

We welcome you and look forward to meeting and speaking with each of you over the next couple of days!

Sincerely,

Ted Slater, Merck, Inc.
CSHALS Conference Chair



Schedule

ROOM ASSIGNMENTS

Tutorial: Charles Suite

Posters, Lunch & Reception: Parkview Room

Wednesday, February 23

7:30 A.M. – 1:00 P.M.	REGISTRATION	
8:30 A.M. – 10:30 A.M.	W3C TUTORIAL	<i>Eric Prud'hommeaux, W3C</i>
		<i>Multi-stakeholder Perspectives on Translational Medicine</i>
10:30 A.M. – 11:00 A.M.	BREAK	
11:00 A.M. – 12:15 P.M.	RPI-LED HANDS-ON TUTORIAL	<i>Timothy Lebo & Jim McCusker, RPI</i>
		<i>Semantic Healthcare and Life Sciences Tutorial: mashing HC and LS Data</i>
12:15 P.M. – 1:00 P.M.	LUNCH	
1:00 P.M. – 3:00 P.M.	RPI-LED HANDS-ON TUTORIAL	<i>(continues)</i>
3:00 P.M. – 3:15 P.M.	BREAK	
3:15 P.M. – 5:00 P.M.	RPI-LED HANDS-ON TUTORIAL	<i>(continues)</i>
4:00 P.M. – 7:00 P.M.	REGISTRATION	
4:00 P.M. – 5:00 P.M.	POSTER (AUTHOR) SET-UP	
5:00 P.M. – 7:00 P.M.	POSTER RECEPTION	

SPONSOR HIGHLIGHTS

Merck & Co., Inc. is a global research-driven pharmaceutical company dedicated to putting patients first. Established in 1891, Merck discovers, develops, manufactures and markets vaccines and medicines to address unmet medical needs. The company devotes extensive efforts to increase access to medicines through far-reaching programs that not only donate Merck medicines but help deliver them to the people who need them. Merck also publishes unbiased health information as a not-for-profit service





SCHEDULE

ROOM ASSIGNMENTS

Posters, Lunch & Reception: Parkview Room

Keynote, Speaker & Tech Talk Presentations: Charles Suite

Thursday, February 24

7:30 A.M. – 10:00 A.M.	REGISTRATION	
7:30 A.M. – 8:30 A.M.	CONTINENTAL BREAKFAST	
9:00 A.M. – 9:15 A.M.	WELCOME & OVERVIEW	<i>Ted Slater, Conference Chair</i>
9:15 A.M. – 10:00 A.M.	KEYNOTE 1	<i>Toby Segaran</i>
	<i>How to Argue for Semantics</i>	
10:05 A.M. – 10:30 A.M.	BIOLOGICS, COMPOUNDS & CHEMISTRY	<i>Christopher Baker</i>
	<i>Semantic Infrastructure for Automated Small Molecule Classification and Data Mining for Lipidomics</i>	
10:30 A.M. – 10:45 A.M.	BREAK	Coffee Break sponsored by BIOGEN IDEC
10:45 A.M. – 11:10 A.M.	BIOMOLECULAR SEMANTICS	<i>James McCusker</i>
	<i>Conceptual Interoperability and Biomedical Data</i>	
11:15 A.M. – 11:40 A.M.	BIOMOLECULAR SEMANTICS	<i>Dexter Pratt</i>
	<i>BEL (Biological Expression Language): Using Causal Relationships to Represent Scientific Findings in Molecular Biology in Support of Applications</i>	
11:45 A.M. – 12:10 P.M.	NEW & INNOVATIVE	<i>Martin Romacker</i>
	<i>Semantic Representation of Events in the Pharmaceutical Industry</i>	
12:15 P.M. – 12:25 P.M.	TECH TALK 1	<i>Jan Aasman, Franz Inc.</i>
	<i>RDF Browser for Pharma Discovery and Visual Query Building</i>	
12:30 P.M. – 12:40 P.M.	TECH TALK 2	<i>Vishal Gupta, Elsevier & SciVerse Platform Embraces Semantic Applications</i>
	<i>Ari Tuchman, Quantified</i>	
12:40 P.M. – 1:30 P.M.	LUNCH	
1:30 P.M. – 2:15 P.M.	KEYNOTE 2	<i>Lawrence Hunter,</i>
	<i>Computational Acceleration of Biomedical Discovery</i>	
2:20 P.M. – 2:45 P.M.	SAFETY, EFFICACY & OUTCOMES	<i>Sherri Matis-Mitchell</i>
	<i>PharmaConnect: Development of an Integrated Knowledge Platform by Extracting, Integrating and Analyzing Information to Support Systematic, Evidence Based Decision Making in R&D</i>	
2:50 P.M. – 3:15 P.M.	SAFETY, EFFICACY & OUTCOMES	<i>Vicki Seyfert-Margolis</i>
	<i>Advancing Regulatory Science for Public Health — An FDA Perspective</i>	
3:20 P.M. – 3:45 P.M.	CLINICAL HARMONIZATION	<i>Eric Neumann</i>
	<i>Semantic Analysis and Visualization of Clinical Data</i>	
3:45 P.M. – 4:00 P.M.	BREAK	
4:00 P.M. – 4:25 P.M.	GENOMICS & GENETICS	<i>James Balhoff</i>
	<i>The Phenoscape Knowledgebase: Linking Evolutionary Diversity to Genetic Data Using Phenotype Ontologies</i>	
4:30 P.M. – 4:55 P.M.	NEW & INNOVATIVE	<i>Therese Vachon</i>
	<i>Fueling Knowledge Federation Using Terminological Services</i>	
5:00 P.M. – 5:45 P.M.	KEYNOTE 3	<i>Charles Mead</i>
	<i>Next-generation Architecture for caBIG</i>	
5:45 P.M. – 5:50 P.M.	DAILY CLOSING REMARKS	<i>Ted Slater, Conference Chair</i>



SCHEDULE

ROOM ASSIGNMENTS

Posters, Lunch & Reception: Parkview Room

Keynote, Speaker & Tech Talk Presentations: Charles Suite

Friday, February 25

7:30 A.M. – 10:00 A.M.	REGISTRATION	
7:30 A.M. – 8:30 A.M.	CONTINENTAL BREAKFAST	
8:30 A.M. – 8:45 A.M.	REVIEW — PREVIOUS DAY	<i>Ted Slater, Conference Chair</i>
8:45 A.M. – 9:30 A.M.	KEYNOTE 4 <i>Semantics for Computational Workflows: A Top Ten List</i>	<i>Yolanda Gil</i>
9:35 A.M. – 10:00 A.M.	ONTOLOGIES & KNOWLEDGE BASES <i>NoSQL: New Possibilities for Distributed Scientific Data Management, Workflow and Collaboration</i>	<i>Mike Miller</i>
10:05 A.M. – 10:30 A.M.	ONTOLOGIES & KNOWLEDGE BASES <i>NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources</i>	<i>Nigam Shah</i>
10:30 A.M. – 10:45 A.M.	BREAK	
10:45 A.M. – 11:10 A.M.	EMERGING & ESTABLISHED STANDARDS <i>Using SWObjects to Create and Query RDF Views</i>	<i>Eric Prud'hommeaux</i>
11:15 A.M. – 11:40 A.M.	EMERGING & ESTABLISHED STANDARDS <i>Publishers' Content Linked with Bioinformatics Data Resources: Working Towards Brokering Standards in the SESL Pilot Project</i>	<i>Dietrich Rebholz-Schuhmann</i>
11:45 A.M. – 12:10 P.M.	EMERGING & ESTABLISHED STANDARDS <i>Rendering Medical Documents in RDF: Strategies and Gotchas</i>	<i>John Madden</i>
12:15 P.M. – 12:25 P.M.	TECH TALK 3 <i>Analysis of Omics Data Using Reverse Causal Reasoning (RCR) in an Integrated Analysis Environment</i>	<i>Dexter Pratt, Selventa</i>
12:30 P.M. – 12:40 P.M.	TECH TALK 4 <i>A 'Killer App' for Semantic Technologies: Point-and-Click Data Integration Tools Make it Easy to Deliver Targeted Semantic Knowledge Bases</i>	<i>Chuck Rockey, IO Informatics</i>
12:40 P.M. – 1:30 P.M.	LUNCH	
1:30 P.M. – 1:55 P.M.	TRANSLATIONAL MEDICINE <i>Semantic Repository of Genomics Experiments</i>	<i>Sudeshna Das</i>
2:00 P.M. – 2:25 P.M.	SEMANTIC WEB AND PHARMA <i>Semantics-enabled Proactive and Targeted Dissemination of New Medical Knowledge</i>	<i>Lakshmish Ramaswamy</i>
2:30 P.M. – 2:55 P.M.	SEMANTIC WEB AND PHARMA <i>Identifying Unexpected Associations in Integrated Biomedical Data Sets: Novel Navigation, Analysis & Visualization Interaction Patterns for Semantic TripleStores</i>	<i>Chris Bouton</i>
3:00 P.M. – 3:15 P.M.	CONFERENCE CLOSING REMARKS & FUTURE ACTIONS	<i>Ted Slater, Conference Chair</i>
3:15 P.M.	CONFERENCE ENDS	



KEYNOTE 1 • TOBY SEGARAN

Data Magnate, Metaweb Technologies
San Francisco, CA, USA



9:15 A.M. – 10:00 A.M.
Toby Segaran,
Data Magnate
Metaweb Technologies
San Francisco, CA, USA

How to Argue for Semantics

ABSTRACT: Many people who could benefit from the techniques used by the semantic web community remain unaware of the advantages conveyed by graphs, URIs and ontologies. In this talk I'll explore perceptions of the semantic web, the kinds of problems people frequently encounter that can be solved with these techniques, and how to explain semantic technology to the uninitiated.

BIOGRAPHY: Toby Segaran is a software developer and the author of the acclaimed O'Reilly title, *Programming Collective Intelligence*, and two new books *Programming the Semantic Web* and *Beautiful Data*. Formerly the Director of Software Development at Genstruct, Toby is now a Data Magnate at Metaweb technologies where he develops techniques to retrieve, parse and reconcile large public datasets. He loves applying data-mining algorithms to everything ranging from pharmaceutical trials to social networks and online dating.

KEYNOTE 2 • LAWRENCE HUNTER

Director of the Computational Bioscience Program and of the Center for Computational Pharmacology, University of Colorado School of Medicine, Aurora, USA



Computational Acceleration of Biomedical Discovery

ABSTRACT: The profusion of high-throughput instruments and the explosion of new results in the scientific literature, particularly in molecular biomedicine, is both a blessing and a curse to the bench researcher. Even knowledgeable and experienced scientists can benefit from computational tools that help navigate this vast and rapidly evolving terrain. However, effective design and implementation of computational tools that genuinely facilitate the generation of novel and significant scientific insights remains poorly understood. In this talk, I will describe a set of efforts that combines natural language processing for information extraction, graphical network models for semantic data integration, and some novel user interface approaches into a system that has recently played a pivotal role in making a significant biomedical discovery.

BIOGRAPHY: Dr. Lawrence Hunter is the Director of the Computational Bioscience Program and of the Center for Computational Pharmacology at the University of Colorado School of Medicine, and a Professor in the departments of Pharmacology and Computer Science (Boulder). He received his Ph.D. in computer science from Yale University in 1989, and then spent more than 10 years at the National Institutes of Health, ending as the Chief of the Molecular Statistics and Bioinformatics Section at the National Cancer Institute. He inaugurated two of the most important academic bioinformatics conferences, ISMB and PSB, and was the founding President of the International Society for Computational Biology. Dr. Hunter's research interests span a wide range of areas, from cognitive science to rational drug design. His primary focus recently has been the integration of natural language processing, knowledge representation and machine learning techniques and their application to interpreting data generated by high throughput molecular biology.



1:30 PM. –2:15 P.M.

Lawrence Hunter,
Director of the
Computational
Bioscience Program,
and of the Center
for Computational
Pharmacology,
University of Colorado
School of Medicine,
Aurora, USA



KEYNOTE 3 • CHARLES MEAD

National Cancer Institute, Center for Biomedical Informatics and Information Technology (CBIIT), Rockville, USA



5:00 P.M. – 5:45 P.M.

Charles Mead,
National Cancer Institute,
Center for Biomedical
Informatics and Information
Technology (CBIIT),
Rockville, USA

Next-generation Architecture for caBIG

ABSTRACT: The caBIG project now has 6+ years of experience with the challenges and benefits of defining, designing, developing, deploying, and evolving a distributed infrastructure to support collaborative data sharing across the translational medicine continuum. As a direct result of both the successes of the first-generation of caBIG and the increasingly complex requirements of the caBIG stakeholder community around not only data sharing, but also more complex analytical and cross-process behavior coordination, the NCI Center for Biomedical Informatics and Information Technology (NCI CBIIT) has begun working on its next-generation architecture for caBIG. This presentation will focus on a detailed enumeration of the distributed processing requirements for the next-generation of caBIG tools and technologies. It will then discuss the core architecture strategies that have been adopted to satisfy these requirements. Of particular importance is the adoption and adaption of a number of standards to support interoperability and automated decision making including such topics as management of cross-platform service-level security, ad hoc and distributed queries, and computationally assembled workflows. In general, the overarching development strategy for the next-generation of caBIG is the combination of leveraging the experience gained and lessons learned both within the caBIG community over the last 6+ years, as well as in the larger internet community as it moves forward in its development of the Web 2.0 strategies, technologies, and tools.



BIOGRAPHY: Dr. Mead has over 35 years of experience in digital signal processing and algorithm development, complex software systems and architectures, and healthcare and life sciences informatics. Dr. Mead has experience in clinical trials methodologies and data management systems, application of the Unified Process, and fundamental healthcare and life sciences informatics issues including terminology management, application of the Health Level Seven (HL7) Reference Information Model (RIM), use of Clinical Data Interchange Standards Consortium (CDISC) standards such as SDTM and ODM, the JANUS data model, and Oracle's HTB development framework. Dr. Mead currently is Chair of the HL7 Architecture Board, past-Chair of the Open Health Tools Architecture Project Team, and a current member of the CDISC Board of Directors.



KEYNOTE 4 • YOLANDA GIL

Associate Director for Research, Intelligent Systems Division, USC/ISI
Research Professor of Computer Science, Information Sciences Institute
University of Southern California, Los Angeles, USA



8:45 A.M. – 9:30 A.M.
Yolanda Gil,
Associate Director for
Research, Intelligent
Systems Division, USC/ISI
Research Professor of
Computer Science
Information Sciences
Institute, University of
Southern California,
Los Angeles, USA

Semantics for Computational Workflows: A Top Ten List

ABSTRACT: In the coming decades, computational experimentation will push the boundaries of current science infrastructure in terms of interdisciplinary scope and integrative models of the phenomena under study. A key emerging concept is computational workflows, which provide a declarative representation of complex scientific applications in terms of the interrelated data retrieval and processing tasks and their mapping to the underlying computational environment. In this talk, I will give an overview of the benefits of using workflows for scientific data analysis, including the management of distributed computations, provenance recording, and reproducibility. I will introduce semantic workflows, which exploit a variety of metadata about data characteristics and data processing algorithms to assist users with significantly more complex analytical tasks. Semantic workflows enable new capabilities for automated workflow generation, reuse, validation, and experiment design that have the potential to increase scientific productivity by orders of magnitude. I will conclude with an overview of the research challenges that lie ahead and the broader benefits of having semantic workflows more widely adopted.



BIOGRAPHY: Dr. Yolanda Gil is Director of Knowledge Technologies at the Information Sciences Institute of the University of Southern California, and Research Professor in the Computer Science Department. She received her M.S. and Ph. D. degrees in Computer Science from Carnegie Mellon University. Dr. Gil leads a group that conducts research on various aspects of Interactive Knowledge Capture. Her research interests include intelligent user interfaces, knowledge-rich problem solving, scientific and grid computing, and the semantic web. An area of recent interest is large-scale distributed data analysis through semantic workflows. Dr. Gil was elected to the Council of the American Association of Artificial Intelligence (AAAI) in 2003, and was program co-chair of the AAAI conference in 2006. She served in the Advisory Committee of the Computer Science and Engineering Directorate of the National Science Foundation. Dr Gil was recently elected chair of the Association for Computing Machinery's Special Interest Group on Artificial Intelligence (ACM SIGART). She leads the W3C Provenance Group, an effort to chart the state-of-the-art and possible standardization efforts in this area.



TUTORIAL

**WEDNESDAY,
FEBRUARY 23**

8:30 A.M. – 10:30 A.M.

Presenter:

Eric Prud'hommeaux, W3C

Multi-stakeholder Perspectives on Translational Medicine Tutorial

Presented by



The W3C's Health Care and Life Sciences Interest Group (HCLS IG) has been working since 2005 with multiple communities and Semantic Web technologies towards goals such as immediate availability of scientific publications; improved synthesis between scientific findings; better patient recruitment for drug trials; and early redirection of non-promising clinical trials.

Today, there is a new convergence of communities in health care and life sciences. Pharmaceutical companies, clinical care providers and individual patients have intersecting interests in Translational Medicine. Pharmaceutical companies have a new interest in more detailed patient records rather than aggregate data because of the shift towards tailored therapeutics.

Consumers advocating for personally controlled health care records are a new audience interested in health care data. Clinicians, pharmaceutical companies and individuals will benefit from health care data which is easy to integrate with genomics, bioinformatics chem informatics and environmental data.

In this tutorial session, we will show you how W3C's HCLS IG is integrating data across these domains. We will demonstrate the use of commodity Semantic Web tools to ask valuable questions of a corpus of health care data, and show how this corpus draws on such systems as the Indivo EHR system, the I2B2 clinical information exchange protocols, and databases backing conventional clinical data stores. Attendees will learn to use and customize these open source tools to meet their clinical or research needs.

TUTORIAL



Semantic Healthcare and Life Sciences Tutorial: mashing HC and LS Data

Presented by



Rensselaer

**WEDNESDAY,
FEBRUARY 23**

11:00 A.M. – 5:00 P.M.

Presenters:

Timothy Lebo and
Jim McCusker, RPI

CSHALS has always been about the practical application of semantic technology to life sciences and pharmaceutical R&D. In keeping with that spirit, this hands-on tutorial will give participants practical experience using Semantic Web tools and technologies to develop mashups using data from the Linked Open Data cloud together with semantic data they create themselves from raw data. Participants will load data into a triple store, query it using SPARQL, use inference to expand the experimental knowledge, and build dynamic visualizations from their results. The tutorial will be loosely based on the book *Programming the Semantic Web**, by Toby Segaran (one of our keynote speakers), Colin Evans, & Jamie Taylor.

AGENDA

11:00 A.M. – 12:15 P.M.

1. PART I – SEMANTIC DATA

Brief introduction to the basics of the Semantic Web, its principles, and its technologies. Here we will discuss what the Semantic Web is, why we need it, what technologies make up the Semantic Web, and how we can use them to link and query data.

- 1.1. Why do we need a Semantic Web?
- 1.2. Intro to basic Semantic Web technology
- 1.3. Just Enough RDF — Converting data to RDF — a few technologies that can be used to convert data to RDF
- 1.4. Data Linking 1.5. SPARQL

1:00 – 3:00 P.M.

2. PART II — MASHUP WORKFLOWS

In this section we will discuss the workflow of building a mashup using semantic data. We will discuss the iterative process of discovering, exploring, linking our data, and publishing on the web. The aim is to learn the semantics and syntax of the data, and how to bring them together.

- 2.1. Data identification — what data sets to utilize
- 2.2. Data understanding — the semantics and syntax of your data
- 2.3. Data linking — how to link together your data

3:15 – 5:00 P.M.

3. PART III — DATA VISUALIZATION

Once we've gone through the data and discovered what's inside it and how to link it to other datasets, we can begin to query the data and visualize the results. This allows us to see new patterns and correlations that weren't visible before. We can also use our visualization to communicate these patterns and ideas to others. We will also cover some popular visualization APIs that can be used for visualization.

3.1 Uses

- 3.1.1. Data discovery — exploring your data through visualization. See new patterns that emerge from visually exploring your data
- 3.1.2. Data communication — communicate the story that is within your data, and see more effective ways to visually present your data for better communication

3.2 Visualization APIs

- 3.2.1. Google visualization APL
- 3.2.2. MIT Simile Exhibit



**THURSDAY,
FEBRUARY 24**
9:15 A.M. – 10:00 A.M.

KEYNOTE 1 TOBY SEGARAN

How to Argue for Semantics

See page 8 for details.

**THURSDAY,
FEBRUARY 24**
10:05 A.M. – 10:30 A.M.

BIOLOGICS, COMPOUNDS & CHEMISTRY

**Semantic Infrastructure for Automated
Small Molecule Classification and Data
Mining for Lipidomics**

Christopher Baker, University of New Brunswick, Saint John, CA

ABSTRACT: Background The development of high-throughput experimentation and combinatorial chemistry has led to astronomical growth in biologically relevant lipids and lipid derivatives identified, screened, and deposited in numerous online databases. At the same time, efforts to annotate, classify, analyze, and link these chemical entities to disparate data sources have largely remained in the hands of human curators using manual or semi-automated protocols. Since chemical function is often closely linked to its structure, and concomitantly, position within a chemical ontology, the accurate classification and annotation of chemical entities is of primary importance in understanding their functionality as well as the full spectrum of potential applications. Unfortunately, neither the expressivity of formal ontologies, nor the potential of Semantic Web Technologies (SWT) to integrate disparate computational services have been fully exploited within the lipidomics and metabolomics communities. Results As part of a case study in the utility of SWT for chemical classification, we have developed a prototype framework for automated lipid classification and annotation. This framework comprises of the following components; Firstly a

formal lipid ontology developed in OWL-DL, which is based in part on the lipid class hierarchy from the LIPIDMAPS database and relevant literature. The Lipid Ontology, [ICBO2009], relies on structural features of small molecules to formally describe lipid classes. Secondly a set of federated Semantic Web services deployed within the SADI framework is used to invoke the automated logical classification task. The first service, a structural annotation service, detects and enumerates relevant chemical subgraphs on a given input chemical graph. Secondly a classifier service assigns chemical entities to appropriate ontology classes by reasoning over class description in the ontology and checking them against the set of chemical subgroups provided by the structure annotation service. We illustrate the utility of these core services using the use case of Eicosanoid classification and combine them with additional SADI services linking the annotated lipids to related proteins found in the biomedical literature or within the public databases. Using these services we further contrast the performance of automated Eicosanoid classification with the existing lipid nomenclature systems and curated lipid databases and reflect on the contribution of our methodology in the context of high-throughput Lipidomics. Conclusions The prototype semantic web service framework we have developed is capable of accurate automatic classification of lipids and integration of information on given chemical entities from relevant databases. The services we provide within this framework can also be reused within other contexts and adapted to diverse lipidomics computational workflows. We conclude that SWT can provide an accurate and versatile means of classification and annotation of chemical entities.



BIOMOLECULAR SEMANTICS

**THURSDAY,
FEBRUARY 24**

10:45 A.M. – 11:10 A.M.

Conceptual Interoperability and Biomedical Data

James McCusker, Rensselaer Polytechnic Institute, Troy, USA

ABSTRACT: Computable semantic interoperability among domain models in biomedicine, as well as interoperability with cross cutting models, has become a major concern in biomedical research. The National Cancer Institute Center for Biomedical Informatics and Information Technology (NCI CBIIT) has begun the next phase for developing caBIG semantic interoperability through the adoption of layered semantics and data models. We discuss a possible mapping between conceptual and logical models. This mapping technique leverages OWL annotation capabilities paired with SKOS representations of existing biomedical ontologies. We show how this technique might provide interoperability among domain and cross-cutting models in caBIG and in other semantic environments. We demonstrate three capabilities that this mapping provides: conversion between domain models and cross-cutting models, conversion between domain models, and domain model-agnostic queries across multiple models. We discuss the application of this technique to the existing caBIG semantics, the proposed caBIG semantics, and to interoperability of biomedical data through the proposed translational research provenance vision.



**BEL (Biological Expression Language):
Using Causal Relationships to Represent
Scientific Findings in Molecular Biology
in Support of Applications**

Dexter Pratt, Selventa, Cambridge, USA

ABSTRACT: The intent of scientific publication is to share knowledge. To do this effectively, scientific documents should be accessible to semantically enabled applications, with critical information encoded in a computationally accessible knowledge representation. This presentation describes the knowledge representation language BEL (Biological Expression Language), a language designed to pragmatically represent scientific findings in molecular biology as causal relationships. BEL was designed to capture knowledge about biological scientific findings as well as their contexts in a user friendly, intuitive way. Findings can be encoded via the representation of experimentally demonstrated causal relationships which are further annotated with information describing biological context, experimental methodology, literature source and curation process. Biological models appropriate for a given analysis or application can be created in a knowledge assembly process in which BEL-encoded findings are integrated, selected, transformed and augmented by inference. Knowledge can be selected based on provenance and biological context information associated with each finding, enabling a strategy where knowledge capture can be well separated from the design of useful models. Each relationship in

**THURSDAY,
FEBRUARY 24**
11:15 A.M. – 11:40 A.M.



a BEL-derived model can be justified by reference to its supporting findings. BEL closely links the represented knowledge to measurable quantities by focusing the ontology on terms denoting abundances and activities of entities at the molecular scale, facilitating the use of BEL-derived models in the interpretation of experimental data sets. BEL terms can be defined by reference to external vocabularies or ontologies, thereby supporting the integration of knowledge from multiple sources. Following eight years of development and proprietary use, BEL has proven to be an intuitive and effective language for scientists, supporting the creation of a large knowledgebase used in the interpretation of 'omics data sets via causal relationship-based analytics. BEL and supporting tools are now being made publicly available to the research community through the introduction of the BEL Web Portal™. The BEL Web Portal™ provides public access to BEL language specifications, documentation, knowledge representation examples, and BEL software tools.

NEW & INNOVATIVE



Semantic Representation of Events in the Pharmaceutical Industry

Martin Romacker, Samuel Läubli & Marc Bux,
Novartis Pharma AG, NIBR-IT, Basel, CH

ABSTRACT: Data feeds from commercial content providers contain information highly relevant to pharmaceutical research. Processing and normalizing the data for in-depth analysis plays an important role in areas like competitive intelligence, strategic alliances or modeling and simulation. Unfortunately, the data is not easy to be integrated and to be semantically syndicated. The standard transfer mode of knowledge in terms of XML files clearly lacks semantics. Additionally, many facts are locked in natural language statements instead of being accessible in a machine-readable and semantically valid representation. The challenge is even larger when content needs to be combined from different feeds. Heterogeneous ways of naming, different semantic typing and different content structures prevent the users from fully exploiting the rich knowledge contained in the feeds. At NIBR-IT, we have implemented an automatic pipeline to process and normalize company names. At the same time, we have created a NLP pipeline which is able to derive facts from statements around phase transitions, mergers and acquisitions or licensing events. By doing so, we transform natural language statement into a normalized semantic representation which uses a Neo-Davidson-like form of notation. The different types of events are captured in a high-level ontology around the event types using OWL.

**THURSDAY,
FEBRUARY 24**

11:45 A.M. – 12:10 P.M.



Having this kind of representation it is now possible to ask queries like “What are the licensing events where Novartis gave a license to any company?”. The company centric events are complemented by knowledge around indications, products and other semantic types. A secondary aspect of this project is to be able to demonstrate to the content providers that it might be an interesting idea to change to a semantically richer and computer-accessible way to deliver data. In the presentation, we will first outline the business rationale behind our project. In the second part, we will give an overview on our way to process and normalize the free text sentences and will explain the Semantic Web approach we have taken to represent data.

TECH TALK 1



RDF Browser for Pharma Discovery and Visual Query Building

Jans Aasman, Franz Inc., Oakland, USA

ABSTRACT: The free-form nature of RDF triplestores offers a lot of flexibility for constructing databases, but that freedom can also make it less obvious how to find arbitrary data for retrieval, error-checking, or general browsing. We will present a graphical triplestore browser that makes data retrieval more pleasant and powerful with a variety of tools for laying out cyclical graphs, displaying tables of properties, managing queries, and building SPARQL queries as visual diagrams.

**THURSDAY,
FEBRUARY 24**

12:15 P.M. – 12:25 P.M.



TECH TALK 2

**THURSDAY,
FEBRUARY 24**

12:30 P.M. – 12:40 P.M.

SciVerse Platform Embraces Semantic Applications

Vishal Gupta, Elsevier, New York, USA; Ari Tuchman, Quantified

ABSTRACT: With an exponential growth of information and wider distribution of services and data sources; integrated search and intelligent semantic discovery become crucial to the success of researchers. SciVerse Applications is an innovative marketplace for applications that enhance the search capacity of researchers by adding semantic search functionality, text and data mining annotation to visualize data and relationships, or integrate data with ontologies and repositories. Developers and researchers can build customized applications that target specific researcher interests and workflows, and drive innovation on the SciVerse platform. Using the SciVerse Framework, third party developers can build applications that appear alongside full text articles, abstracts and search results within the SciVerse product suite. Developers can also integrate external web services and APIs in their applications to mashup with the SciVerse APIs. In this session we will showcase the SciVerse platform and demonstrate two applications driven by semantic technology and developed in collaboration with our partners at Stanford University and Quantifind. ODiSSea semantically expands your query with standard ontologies and public data resources. Quantifind extracts and aggregates data from the corpus associated with a user query and visualizes the associated data and trends.



KEYNOTE 2 LAWRENCE HUNTER

**Computational Acceleration of
Biomedical Discovery**

See page 9 for details.

**THURSDAY,
FEBRUARY 24**

1:30 P.M. – 2:15 P.M.

SAFETY, EFFICACY & OUTCOMES

**PharmaConnect: Development of an
Integrated Knowledge Platform by
Extracting, Integrating and Analyzing
Information to Support Systematic,
Evidence Based Decision Making in R&D**

Sherri Matis-Mitchell, AstraZeneca Pharmaceuticals, Wilmington, USA

ABSTRACT: The Knowledge Engineering initiative within AstraZeneca has recently delivered the first version of a knowledgebase that integrates internal and external evidence for connections between key concepts such as targets, pathways, compounds, diseases, preclinical, and clinical outcome from Chemistry, Competitive, Disease and Safety Intelligence workstreams. This talk will describe the system, architecture, and its development; demonstrate the impact of this new platform with specific examples; and discuss lessons learned during its development. We will also detail linkages to additional data sources and system as well as plans for the future.

**THURSDAY,
FEBRUARY 24**

2:20 P.M. – 2:45 P.M.



SAFETY, EFFICACY & OUTCOMES

**THURSDAY,
FEBRUARY 24**

2:50 P.M. – 3:15 P.M.

Advancing Regulatory Science for Public Health — An FDA Perspective

Vicki Seyfert-Margolis, US Food and Drug Administration, Silver Spring, USA

ABSTRACT: For breakthroughs in science and technology to reach their full potential, FDA must play an increasingly integral role as an agency not just dedicated to ensuring safe and effective products, but also to promote public health and participate more actively in the scientific research enterprise directed towards new treatments and interventions. We must also modernize our evaluation and approval processes to ensure that innovative products reach the patients who need them, when they need them. These new scientific tools, technologies, and approaches form the bridge to critical 21st century advances in public health. They form what we call regulatory science: the science of developing new tools, standards and approaches to assess the safety, efficacy, quality and performance of FDA-regulated products.

CLINICAL HARMONIZATION



Semantic Analysis and Visualization of Clinical Data

Eric Neumann, Clinical Semantics, Bedford, USA

ABSTRACT: Biomedical data generation is continuously growing both in terms of size and complexity. Clinical Study data is complicated by the fact that new forms of associated data are continuously created as technologies emerge, including biomarkers, pathway (mechanistic) knowledge, assay platforms, and model systems. W3C semantic standards such as RDF and OWL have been around for several years, but most informatics specialists are unsure where they can be applied effectively. Semantically Linked Data (SLD) can significantly change the organization and re-use of data without requiring a concomitant investment in data systems. SLD is especially fine-tuned for handling information extracted from literature, and relating it to structured data, even if they exist in other data systems.

**THURSDAY,
FEBRUARY 24**

3:20 P.M. – 3:45 P.M.



GENOMICS & GENETICS

**THURSDAY,
FEBRUARY 24**

4:00 P.M. – 4:25 P.M.

The Phenoscape Knowledgebase: Linking Evolutionary Diversity to Genetic Data Using Phenotype Ontologies

James Balhoff, National Evolutionary Synthesis Center, Durham, USA

ABSTRACT: Objectives and motivation Phenotypic differences among species have long been systematically itemized and described by biologists in the process of investigating phylogenetic relationships and trait evolution. Traditionally, these descriptions have been expressed in natural language within the context of individual journal publications or monographs. As such, this rich store of phenotype data has been largely unavailable for statistical and computational comparisons across studies or integration with other biological knowledge. We have created the Phenoscape Knowledgebase, which consists of a database and web application (<http://kb.phenoscape.org/>). The database combines ontologically annotated phenotypic character data for a large and diverse group of fishes with phenotypic annotations from the ZFIN model organism database. The web application provides query and browsing interfaces which allow users to exploit the the logical framework provide by the ontologies which underpin the data. Method We used OBD (“Ontology-based Database”) to store phenotypic data, from ~50 phylogenetic publications, as statements using terms from ten different OBO ontologies. The phenotypic data, taxa, and specimens in these published data sets were annotated with ontology terms using our curation application, Phenex. In this process free-text phenotype descriptions were converted to semantic representations using an Entity-Quality (EQ) model, combining terms from separate anatomical and qualitative ontologies. The ontologies and annotated



data sets, along with EQ phenotype annotations for zebrafish genes, exported from the ZFIN database, were loaded into OBD using its own triple-based schema. We used the SQL-based OBD reasoner to pre-compute inferred statements and add them to the Knowledgebase. We developed a web services API providing access to the Knowledgebase using the Restlet Java framework. We also developed a Ruby on Rails-based end-user web interface, which allows biologists to query the Knowledgebase, accessing the data via these public web services.

Results The Phenoscape Knowledgebase integrates over 500,000 asserted phenotype statements, concerning ~2500 fish species, with over 20,000 phenotype statements linked to over 3700 zebrafish genes. Users can discover fish species matching arbitrary phenotypic profiles, which can be expressed as queries making use of the hierarchical nature of anatomical, qualitative, and taxonomic ontologies. Moreover, genes influencing these phenotypes can be simultaneously returned. At the same time users can visualize the structure and explore term definitions of the included ontologies. The Knowledgebase has been used to investigate patterns of anatomical coverage within published phylogenetic characters, as well as to generate hypotheses for candidate genes underlying evolutionary losses of both scales and skeletal elements.

Conclusion Ontological annotations of free-text phenotypic data, built with shared community-driven ontologies, constitute a powerful resource when aggregated within a database system which makes full use of the semantic framework provided by those ontologies. For the first time, scientists can search phenotypic content from dozens of phylogenetic publications, querying across anatomical, qualitative, and taxonomic axes.



NEW & INNOVATIVE

**THURSDAY,
FEBRUARY 24**

4:30 P.M. – 4:55 P.M.

Fueling Knowledge Federation Using Terminological Services

Therese Vachon, Novartis Pharma AG, Basel, CH

ABSTRACT: Knowledge proliferation and data silos are well-known buzz words which characterize the way data is produced and stored in the pharmaceutical industry. Most efforts in knowledge mining try to make the knowledge buried in applications and data bases accessible. These efforts are both expensive and tedious. Additionally, not all knowledge can be recovered as the stored information tends to be ambiguous and incomplete. At the Novartis, we have been working on a principled approach to overcome these shortcomings. The basic idea is to create a federation layer based on well controlled terminologies aiming at a uniform wording within and across data repositories. Thus, we have been collecting and defining meaningful atomic units (basic concepts) together with their lexical representations (terms) in a knowledge integration framework. Within that framework we maintain a number of terminologies (like indication, company, target, gene, assay method). The terminologies are organized in terms of taxonomies and complemented by referential knowledge, so-called cross references or pointers which link out to other repositories. One of our objectives is to stay compatible with the major resources of the open biomedical community. With regards to the coverage, our terminologies focus on the terms which are really relevant to research at NIBR. Cross referencing is a powerful but formally simple means to link out to other knowledge repositories to get access to additional information. Our methods to maintain and enhance the different terminologies in our framework depend mainly on the concept type. We have different levels of automatic generation, versus intellectual curation of the content related to indications, companies or genes. We believe that for each of these concept types



there as an optimal balance between automation and curation – the former being prone to errors and the latter being time consuming and therefore expensive. Furthermore, we intend to make the maintenance process more and more a collaborative task where scientists can access, review and modify the content according to their role profile. An important success factor for the widespread usage of terminologies is to bring them seamlessly to the point of usage. Consequently, we have implemented a service layer providing SOAP and JSON Web Services as well as a REST API. Importantly, the users have access to the knowledge without slowing their work and without having to leave the active application. The increasing usage of these services both in number of applications and in number of calls clearly demonstrates the importance of the flexible integration layer. It is important to mention that for some of the concept types we have reached a critical mass in usage which allows us to run queries across systems or provide concept centric views joining internal and external data. As we can demonstrate the benefits from using our resources more and more people and organizations are starting to buy in. In our oral presentation, we would like to give an overview on our approach to “Terminology Management” and illustrate how we represent knowledge (terminological and referential). Finally, some use cases demonstrate how the services are applied.

KEYNOTE 3 CHARLES MEAD

Next-generation Architecture for caBIG

See page 10 for details.

**FRIDAY,
FEBRUARY 24**
5:00 P.M. – 5:45 P.M.



**FRIDAY,
FEBRUARY 25**
8:45 A.M. – 9:30 A.M.

**FRIDAY,
FEBRUARY 25**
9:35 A.M. – 10:00 A.M.

KEYNOTE 4 YOLANDA GIL

**Semantics for Computational workflows:
A Top Ten List**

See page 12 for details.

ONTOLOGIES & KNOWLEDGE BASES

**oSQL: New Possibilities for Distributed
Scientific Data Management, Workflow
and Collaboration**

Mike Miller, Cloudant Inc., Boston, USA

ABSTRACT: Inspired by new problems (exploding sensor data, complex workflows, geo-distribution, etc.), there has been a dramatic renaissance of alternatives to classic relational database management systems. We briefly review these “NoSQL” implementations including key/value stores, big tables, document stores and graph stores. Next we focus on specific qualities that enable new possibilities for scientific data management, processing and analysis, in particular: flexibility, scalability, expressiveness, REST interfaces, concurrency, replication and cloud hosting. Finally, we discuss relevant applications in physical and biological sciences.



NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources

Nigam H. Shah, Stanford School of Medicine, California, USA

ABSTRACT: The volume of publicly available data in biomedicine is constantly increasing. However, this data is stored in different formats on different platforms. Integrating this data will enable us to facilitate the pace of medical discoveries by providing scientists with a unified view of this diverse information. Under the auspices of the National Center for Biomedical Ontology, we have developed the Resource Index—a growing, large-scale index of more than twenty diverse biomedical resources. The resources include heterogeneous data from a variety of repositories maintained by different researchers from around the world. Furthermore, we use a set of 200 publicly available ontologies, also contributed by researchers in various domains, to annotate and to aggregate these descriptions. We use the semantics that the ontologies encode, such as different properties of classes, the class hierarchies, and the mappings between ontologies in order to improve the search experience for the Resource Index user. Our user interface enables scientists to search the multiple resources quickly and efficiently using domain terms, without even being aware that there is semantics under the hood.

**FRIDAY,
FEBRUARY 25
10:05 A.M. – 10:30 A.M.**



EMERGING & ESTABLISHED STANDARDS

**FRIDAY,
FEBRUARY 25**

10:45 A.M. – 11:10 A.M.

Using SWObjects to Create and Query RDF Views

Eric Prud'hommeaux, W3C, Cambridge, USA

ABSTRACT: SPARQL CONSTRUCTs, RIF and other rule forms allow us to trivially tailor views of RDF data from sources like turtle files, GRDDL'd XML documents, RDF databases, or conventional relational databases. Views over databases are especially practical if they can be virtual, that is, SPARQL queries over the virtual graph are mechanically transformed into SPARQL queries for RDF databases or SQL queries for conventional relational data (e.g. and Employees table and an Address table). This talk will discuss the utility of such an architecture, including efficient access to RDBs and pipelines of transformation services supported by parties other than the custodians of the final data resources. Real-world examples will include using the SWObjects toolbox to view Gene Ontology as BioPAX and to ask questions which unify across Uniprot and GO databases.



**Publishers' Content Linked with
Bioinformatics Data Resources: Working
Towards Brokering Standards in the SESL
Pilot Project**

Dietrich Rebholz-Schuhmann, European Bioinformatics Institute,
Hinxton, UK

ABSTRACT: The SESL pilot project explores the technical feasibility for federated querying across full text literature and bioinformatics databases. Five Life Science and Pharmaceutical companies have collaborated with four publishers and the Rebholz group (EMBL-EBI) to extract selected data from bioinformatics databases (Uniprot, OMIM and ArrayExpress) and full text literature with focus on human diseases related to Type 2 diabetes mellitus. Gene to disease related assertions have been delivered through a single point of query to the scientist users. The pilot implements the integration of content from public resources and extracted information from the scientific literature into a shared infrastructure based on Semantic Web technology. The SPARQL endpoint is hosted at the EBI and can be accessed remotely through SPARQL queries, a Web browser based graphical user interface or through a SOAP Web services client. The project delivers a preliminary set of standards describing the minimal infrastructure necessary to support a biology brokering service and the provision of a prototype instance of that infrastructure as a public demonstrator.

**FRIDAY,
FEBRUARY 25**
11:15 A.M. – 11:40 A.M.



EMERGING & ESTABLISHED STANDARDS

**FRIDAY,
FEBRUARY 25**

11:45 A.M. – 12:10 P.M.

Rendering Medical Documents in RDF: Strategies and Gotchas

John F. Madden, Duke University, Durham, NC, USA

ABSTRACT: Clinical medical records consist of documents such as laboratory reports, physician’s progress notes, admission summaries, etc.. They often contain a mixture of full-sentence, natural language text and “bullet-point” or form-like content. Non-explicit knowledge (the document’s purpose, genre, temporal context, author’s background knowledge, etc.) as well as references to external assertions found in other documents heavily condition the meaning of such documents. Rendering the content of such a document in RDF is a complex act of interpretation, akin to translation. There is no single “correct” RDF rendering. We will examine sample medical documents and study some possible renderings into RDF/OWL, with the purpose of highlighting common challenges including the following:

- dealing with anaphora, i.e., candidate triples whose appropriate subject is ambiguous or multiple (“sodium 142 mM:: Whose sodium? The patient’s? The patient’s serum? The sample of the patient’s serum delivered to the laboratory? etc.?”)
- instances versus classes, especially when using legacy vocabularies (“Jim has influenza”: Does Jim have SNOMED-influenza, or does he have an instance of SNOMED-influenza?)
- dealing with references to assertions in other documents (“My colleague Dr. Smith diagnosed pneumonia last week”: Is pneumonia the relevant fact, or is the diagnosis of pneumonia the relevant fact? How do I represent the difference?)



Analysis of Omics Data Using Reverse Causal Reasoning (RCR) in an Integrated Analysis Environment

**FRIDAY,
FEBRUARY 25
12:15 P.M. – 12:25 P.M.**

Dexter Pratt, Selventa, Cambridge, USA

ABSTRACT: A critical challenge in modern biology is the development of informative mechanistic hypotheses from a large amount of information produced by today’s molecular profiling methods, such as microarray and next-generation sequencing methods. To address this challenge, Selventa™ utilizes the Genstruct® Technology Platform (GTP). In this talk, we present Reverse Causal Reasoning (RCR), one of the key analytic capabilities of the GTP, to translate large-scale data into meaningful, testable molecular mechanisms. RCR is an automated reasoning technique that processes networks of causal relationships to formulate hypotheses, and then evaluates those hypotheses against data sets of differential measurements. RCR analysis attempts to answer the question “What signaling differences could lead to the observed differences in measured quantities?” The method is called “reverse” because the evaluation of each hypothesis effectively reasons from observed effects to identify potential causes. RCR is applicable to any molecular profiling data type, including microarray, RNA-Seq, proteomic/phosphoproteomic, and metabolomic data. The output of RCR is a set of mechanistic hypotheses that represent upstream controllers of significant differential measurements, ranked by the calculation of statistical figures of merit for relevance and accuracy. These hypotheses can be assembled into causal chains and larger networks to interpret the data set at the level of interconnected mechanisms and processes. Implementation of RCR uses a large knowledgebase encoded using BEL (Biological Expression Language) based on manual curation of the scientific literature. RCR methodology is available through a web-accessible investigation suite that includes facilities for managing experimental data, performing RCR and analyzing hypotheses generated by RCR.



TECH TALK 4

**FRIDAY,
FEBRUARY 25**

12:30 P.M. – 12:40 P.M.

A ‘Killer App’ for Semantic Technologies: Point-and-Click Data Integration Tools Make it Easy to Deliver Targeted Semantic Knowledge Bases

Chuck Rockey, IO Informatics, Berkeley, USA

ABSTRACT: HCLS present unusually dynamic needs for data integration, application and search workflows. Semantic technologies are uniquely suited to provide game-changing integration tools for HCLS through their far more flexible methodologies for data integration and knowledge building.

Traditional data warehousing and federation approaches often choke on the heterogeneity of data sets involved in HCLS (e.g., multiple experimental data silos, public data sets, and clinical data as well as NLP results from scientific, competitive and patent literature.) Even if projects enjoy initial successes, traditional solutions quickly start to break down as novel application requirements and data sets are added.

This talk will describe and demonstrate semantic methods that make it possible for domain experts as well as ontologists and informatics experts to quickly build, modify and extend integrated knowledge bases. Formal ontologies or application ontologies are assembled on the fly as part of the integration process — all without programming or hand-coding of RDF. We’ll show you how, in 10 minutes.



Semantic Repository of Genomics Experiments

Sudeshna Das, Massachusetts General Hospital,
Harvard Medical School, Cambridge, USA

ABSTRACT: Objectives Genome-wide experiments are routinely conducted to study gene expression, DNA-protein binding and epigenetic status. The importance of structured meta-data for these experiments for integration and reuse is widely recognized. For this purpose, first the MIAME standard was developed for microarrays and recently the ISA-TAB format was published as a generalized format for experiments employing omics technologies. Several MIAME-compliant repositories exist for genomics data, notably Array Express and GEO. However, these are not yet widely available as Linked Data compliant with standard biomedical ontologies such as the MAGE Ontology (MO), the Ontology for Biomedical Investigators (OBI) and the Experiment Factor Ontology (EFO). Researchers need friendly, useful and reusable software environments that can automatically produce such Linked Data. Method We have developed reusable software to build semantic repositories of genomics experiments. Our software is based on the open source content-management system Drupal (www.drupal.org). The primary content type is an experiment; which has a title, researcher, design details and is comprised of one or more bioassays. The experiment can be linked to publication(s). Bioassays are processes that have biomaterials and technologies as participants and data files as output. The main classes are mapped to MO & OBI. Biomaterials have various characteristics such as organism, disease state and cell types. These characteristics are mapped to existing published biomedical concepts. The data is entered in a structured format – thus, eliminating the need for future curation. We then use RDF

**FRIDAY,
FEBRUARY 25**

1:30 P.M. – 1:55 P.M.



modules in Drupal to produce Linked Data & a SPARQL endpoint. Results We have developed two repositories using this software in separate domains. One of them is a repository for hematopoietic stem cell data (<http://bloodprogram.hsci.harvard.edu>). It contains over 100 microarray, transcription factor binding and histone modification experiments. The majority of the data is from microarray experiments performed on model organisms (mouse and zebra fish) and encompasses various cell types and disease states. The cell types were mapped to the Cell Type (CL) Ontology. The other repository comprises of microarray profiles from Parkinson's disease patients (<http://pdexpression.org/>). The disease subtypes and tissue of subjects were mapped to standard terms with the help of the NCBO (National Center for Biomedical Ontologies) annotation tool. The use of standard terminologies to describe the biomaterials allows interoperability with other repositories. However, finding the most appropriate mappings still remains a challenge. For example, when mapping the cell types – there were quite a few missing entries, whereas “Parkinson's Disease” was found in over 20 systems. Addressing these issues is as much a social process as a technological one. Conclusions The main benefit of our software is the ability to create Linked Data in a synchronous manner that eliminates the need for latter curation. For each domain we can deploy an instance of the software that is pre-populated with relevant terms (mapped to existing terminologies) from that field. As more communities begin to adopt such reusable infrastructure and make Linked Data available, we will begin to address the integration challenge that is currently posed to biomedical researchers.



Semantics-enabled Proactive and Targeted Dissemination of New Medical Knowledge

**FRIDAY,
FEBRUARY 25**

Lakshmish Ramaswamy, University of Georgia, Athens, USA

2:00 P.M. – 2:25 P.M.

ABSTRACT: The body of knowledge in the field of medicine is expanding at a tremendous pace. The number of citations in MEDLINE grew by more than 700,000 in 2009 and it is expected to grow by 1 million this year. This includes discovery of new drugs, previously unknown reactions to existing drugs, and new treatments. Some of the discoveries are so important that they have to reach the end-practitioners quickly so that they can act upon new knowledge, possibly by altering the course of treatment of relevant patients. Typically, medical knowledge dissemination occurs through channels such as conferences, medical journals, and memos. In the past decade, Web has, to some extent revolutionized medical knowledge dissemination by providing advanced search capabilities. However, this mode of knowledge dissemination is passive, and it has significant limitations. First, it requires the doctors to periodically search the online databases, which places additional burden on doctors. Second, the time lag for a doctor to become aware of a research depends upon how often she searches the online databases. Third, even after a doctor becomes aware of certain medical information, it would take additional time for her to search through the patient records to find out the patients to whom the information would be relevant. These limitations highlight the need for a proactive medical information dissemination paradigm. Our vision is to design and develop a semantics-enabled framework for proactive and targeted dissemination of new medical knowledge. We believe that such a system has to achieve two major design goals. First, in order to prevent information overload, information dissemination has to be targeted in the sense that a



doctor should receive alerts about new discoveries if the information is likely to be relevant to one or more of her patients. Second, the additional workload on the doctor for participating in the system should be very minimal. In other words, the system should function based upon the information that is recorded during the examination and treatment of patients. Towards achieving these two goals, our main idea is to utilize patients' electronic medical records (EMRs) to identify information in scientific articles, memos, etc. that are relevant to a particular patient and alert her doctor accordingly. Several research challenges have to be addressed in order to make such a system efficient and scalable including: (1) EMRs, research publications (from PubMed, etc.), and memos from organizations such as CDC and FDA have to be automatically annotated with medical ontology-based semantics-rich metadata; (2) Novel, semantics-driven algorithms for retrieving, filtering and ranking information relevant to a particular EMR have to be designed; (3) In the interest of system scalability techniques to cluster EMRs based upon their semantic-similarity need to be utilized; and (4) The system has to be continuously tuned based upon explicit and implicit feedback from the users to maintain and improve its effectiveness. In this talk, we will motivate this work through real-world examples. We will elaborate upon the above challenges, discuss our ideas towards addressing them, and present an architecture for our framework.



Identifying Unexpected Associations in Integrated Biomedical Data Sets: Novel Navigation, Analysis & Visualization Interaction Patterns for Semantic TripleStores

Christopher Bouton, Entagen, LLC, Newburyport, USA

ABSTRACT: A promise of semantic technologies is the facile integration of large quantities of disparate data. In the biomedical research and development (R&D) sector this type of integration is essential for the potential identification of connections across entity domains (e.g. compound to targets, targets to indications, pathways to indications). However, the vast majority of data currently utilized in biomedical R&D settings is not integrated in ways which make it possible for researchers to intuitively navigate, analyze and visualize these types of interconnections. Using the Linking Open Drug Data (<http://esw.w3.org/HCLSIG/LODD>) data sets, we have been experimenting with novel forms of biomedical data integration, navigation, analysis & visualization through the development of a web-based, rich-internet application (RIA). An essential goal of this work is the creation user interface paradigms which enable “bench” researchers to intuitively identify unexpected associations which may drive their research forward through the iterative process of effective hypothesis generation and subsequent testing.

**FRIDAY,
FEBRUARY 25**
2:30 P.M. – 2:55 P.M.



POSTERS

POSTER 1

Contextual Understanding of Experimental Data Via Formal Semantic Integration of NLP-extracted Content with other Semantically Integrated Resources

Robert Stanley, Jason Eshleman, Erich Gombocz, IO Informatics, Inc., Berkeley, CA, USA; David Milward, Linguamatics Ltd., UK

Biological systems are inherently complex. Experimental results, especially if they cover multiple experimental modalities or diverse biological responses, are difficult to interpret out of context. This is a key area for the application of semantic technologies. The first step is the integration of analytical results under a well-formed application ontology. Extensible semantic integration standards such as RDF, N3 and OWL are used to create triples-based coherent dynamically extensible and remappable data models. This first step supports the rapid creation of coherent experimental correlation networks and provides a statistically relevant view of system perturbations. However, this does not necessarily provide a better understanding of biological functions involved. In order to achieve contextual understanding, these networks need to be further enriched by adding mechanistic knowledge. This contextual understanding requires the ability to bring in resources (either through direct connections or via queries to SPARQL endpoints) relevant to biological functions. The addition of information about interactions, pathways or other information from previous observations is relevant to describe biological processes, which may otherwise be missed. Natural Language Processing (NLP) can be used to extract relationships between concepts from resources such as scientific journal articles, collaborations, comparative studies and clinical trials. When the NLP extracted relationships are semantically integrated with experimental findings,

POSTERS



the consequential view of the biological system is enhanced. Using thesauri to harmonize classes and relationships from those extracts and merging them into a dynamically extended application ontology results in functionally connected experimental results. This approach makes it possible to apply biomarker patterns or molecular signatures derived from the network to answer complex biological questions, and also to apply them actively for screening and decision support. This poster describes a use-case in which multiple experimental datasets (micro-RNA, sequencing, gene expression, drug target assays) have been semantically integrated, enriched with public knowledge resources (tissue-specific gene expression and regulation [TIGER], human RNA drug targets [TargetScan], miRBase, Microcosm, Disasome) and supplemented with NLP extracted relationships concerning specific diseases (in this case, severe renal failure) from a variety of articles and other sources. Tools used in this scenario were IO Informatics' Sentient Knowledge Explorer for the semantic integration of experimental data, ontology import, network visualization and graphical SPARQL queries in conjunction with relationships extracted from MEDLINE abstracts by Linguamatics' I2E enterprise text mining platform. The resulting semantic network provides a reliable qualification of drug targets with broader applications. The kidney-disease related profiles generated in this example are based on contextual understanding of the biological functions involved in the disease and their manifestation in grounded experimental observations as well as through verification with mined content from trusted resources. Such methodology significantly impacts the way life sciences' and drug discovery research is leading towards more effective drugs, and for widespread use in personalized medicine to improve the quality of life.



POSTERS

POSTER 2

DrugTree: A Phylogenetic Platform to Study Protein-ligand Binding Relationships in the Drug Discovery Process

Katherine Herbert, Nina Goodey, D. Jason Seraydarian, Roberto Suarez, Shreya Achar, Department of Computer Science, Montclair State University, Montclair, NJ, USA

The discovery of drugs that have the desired pharmacological profiles is critical for human health and survival yet time consuming and expensive. Consequently we must aim for obtaining maximum benefits from those medicinal compounds that have already been identified and found to have favorable properties. The DrugTree Project creates toolkit for scientists interested in understanding the broader implications of the relationship between phylogenetics and the binding between a homologous set of enzymes and their corresponding ligands and inhibitors. Phylogeny is a useful context in which to view these relationships: As a protein evolves, one feature that changes is the binding pocket and hence binding specificity. Consequently, evolutionary relationships can provide predictive power to establish the binding between a given ligand and a homolog based on known binding relationships within a protein family. Insight into which phylogenetically prevalent amino acid changes within the binding site are responsible for different ligand specificities amongst the homologs in a family may also be gained. The DrugTree Project has completed a prototype World Wide Web-based computing system that integrates both phylogenetic data and analyses about enzymes with known information about their ligands and inhibitors. Currently, no one data repository integrates the drug-target, protein-ligand curated datasets with a large, popular protein

POSTERS



database like UniProt and then gives tools to allow users to view these datasets in a phylogenetic-meaningful context. The DrugTree tool integrates data from UniProt (<http://www.uniprot.org/>), the BindingDB (<http://www.bindingdb.org/>) and BRENDA (<http://www.brenda-enzymes.org/>) databases to allow the user to create trees with data from both UniProt, with its massive non-redundant database and the data from the known inhibitor repositories. The system initially integrates these three dataset, creating a local repository. Via Web interface, it allows a user to create a phylogenetic tree for a homologous set of enzymes. It then enables the user to perform phylogenetic reconstruction analyses via parsimony techniques with a select few phylogenetics reconstruction algorithms. Finally, the tool then maps allows the user to view the compounds that inhibit or bind each homolog next to the enzyme name. This poster introduces the DrugTree tool. It will demonstrate its effectiveness through analysis of a subset of dihydrofolate reductase proteins and some of the set's known inhibitors. Dihydrofolate reductase is both an important target and a good model system: this enzyme has recently been of interest as a drug target in global health issues including treatment of various parasitic diseases such as malaria, African sleeping sickness, Changa's disease, and tuberculosis. Many sequences and crystal structures are available for dihydrofolate reductase and purification is easy due to the commercial availability of affinity chromatography resin specific for this enzyme. Therefore, it is ideal in verifying our results. It will also discuss our future development plans for the DrugTree platform.



Posters

POSTER 3

Data Driven Derivation of Canonical Eligibility Criteria for Clinical Trials

Saranya Krishnamoorthy, Dinakarpanthian Deendayal, Saranya Krishnamoorthy, Yugyung Lee, University of Missouri, Kansas City, Missouri, USA

Recruitment of subjects for clinical trial research is currently an inefficient and time-consuming process in the development of a new drug. Recruitment challenges are particularly difficult for studies involving vulnerable populations, especially those with psychiatric disorders. The other major hurdle to automate the process is that eligibility criteria are written in free text that cannot be reliably parsed or processed computationally. To overcome these obstacles, we created an intelligent online system which targets the following two goals: Helping recruiters to develop/specify a standardized representation of eligibility criteria. Automate selection of candidates for mental health research studies. As proof of concept, the methodology has been developed and validated on a corpus of 701 clinical trials on Generalized-Anxiety-Disorder containing 2765 and 4411 redundant inclusion and exclusion criteria set.

This paper presents a complementary data driven approach to help find a minimal non-redundant representation of an arbitrary collection of clinical trial eligibility criteria and automates the recruitment of patients for clinical trials. Thus our system allows the recruiters to have the flexibility of using free-text while the semantics of the criteria are captured for computer readability. We would like to acknowledge National Institute of Mental Health for funding the project (1R43MH085372-01A1). The complete abstract is available at: <http://www.iscb.org/cshals2011-program/cshals2011-poster-presentations>

Posters



POSTER 4

Integrating Multi-Dimensional Genomic, Proteomic and Clinical Data of Inflammation and Injury

Wenzhong Xiao, Massachusetts General Hospital, Stanford Genome Technology Center, Stanford, CA USA

Recent developments in high throughput technologies have enabled direct studies of patients' genomic response to diseases and treatments, and new computational methods need to be developed to translate the large amount of genomic, proteomic and clinical data to new knowledge in medicine. Over the past nine years, the Inflammation and the Host Response to Injury Glue Grant Consortium has utilized multiple experimental tools to study the temporal immune-inflammatory response in blood leukocytes and sub-populations from over 500 severely injured patients, together with their comprehensive clinical information. We are developing semantics-based approaches in integrating these genomic and proteomic data with clinical information of patients to elucidate disease mechanism and predict patient outcomes.



POSTERS

POSTER 5

Translational Medicine in Action: Linking and Visualizing a Network of Biomedical Research Scientists using Nexus

Janos Hajagos, Erich Bremer, Janos Hajagos, Tammy Diprima, Stony Brook University School of Medicine, Stony Brook, NY, USA

The goal of translational medicine is to translate basic science research into advances in clinical medicine. One way to meet this goal is to pair up basic scientists with clinical researchers who share common research interests. The challenge is that the terms used by each group do not perfectly align. To demonstrate the utility of using semantic web technology in translational medicine we apply it to interconnect clinical and basic scientists research interests. The research interests of SUNY Reach faculty were obtained from MeSH terms of publication data and are expressed in the VIVO ontology normalized to the UMLS. The VIVO ontology is part of the NIH funded VIVO project to interlink research scientists across different institutions. To explore novel interconnections in the network of research scientists the Nexus visualization environment was utilized. Nexus, a locally developed project, is a semantic web visualization tool built on the OpenSimulator platform. Nexus allows collaborative real time viewing and annotating of RDF data in a 3D environment.

19th Annual International Conference on Intelligent Systems for Molecular Biology
& 10th European Conference on Computational Biology



KEYNOTE SPEAKERS



**2011 ISCB ACCOMPLISHMENT
BY A SENIOR SCIENTIST AWARD**

Michael Ashburner
University of Cambridge, UK



2011 ISCB OVERTON PRIZE

Olga Troyanskaya
Princeton University, USA



ISCB FELLOW KEYNOTE

Alfonso Valencia
*Spanish National Cancer
Research Centre (CNIO), Spain*



Bonnie Berger
*Massachusetts Institute
of Technology, USA*



Luis Serrano
*Centre for Genomic
Regulation, Spain*



Janet Thornton
*European Bioinformatics
Institute, Cambridge, UK*

CONFERENCE CHAIRS

Burkhard Rost, *Conference Chair,*
Technical University Munich, Germany

Michal Linial, *Conference Vice-chair,*
The Hebrew University of Jerusalem, Israel

Peter Schuster,
Conference Honorary Co-chair,
University of Vienna
(Professor Emeritus), Austria

Kurt Zatloukal,
Conference Honorary Co-chair,
*Genome-Austria Tissue Bank, Medical
University of Graz, Austria*

REGISTER EARLY AND WIN!

Register by May 15 & enter to win
FOUR NIGHTS ACCOMMODATION
at the Hilton Vienna
for your conference stay!



An Official Conference of
the International Society
for Computational Biology

www.iscb.org/ismbecb2011



An official conference of the
International Society for
Computational Biology