



**Conference on
Semantics in
Healthcare
And
Life
Sciences**

February 25-27, 2009
MARRIOTT BOSTON CAMBRIDGE
CAMBRIDGE/BOSTON, USA





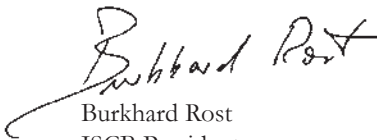
Dear C-SHALS participant,

Welcome to the ISCB Conference on Semantics in Healthcare and Life Sciences, brought to you with the generous support of our sponsors: IO-Informatics, Merck and Pfizer.

The C-SHALS conference focuses on pharmaceutical applications of semantic technologies, and provides a uniquely intimate forum for discussion among visionaries, leaders and those using or looking to use intelligent information technologies in pharmaceutical R&D. For their commitment to developing an exceptional line up of speakers and program topics I would like to personally acknowledge conference chair Eric Neumann and organizing committee members Mike Bevil, Joanne Lucianno, Ted Slater and Susie Stephens. ISCB's Director of Conferences, Steven Leard, also worked his natural brand of magic, together with BJ Morrison McKay, ISCB Executive Officer, to translate the organizing committee's vision into a cohesive program that fosters discussion and information-sharing among all attendees. ISCB is dedicated to providing high quality meetings, and I believe each of these individuals have helped to achieve that goal with this second annual C-SHALS event. ISCB has been particularly pleased to once again feature a pre-conference tutorial presented by W3C's Semantic Web for Health Care and Life Sciences Interest Group. We value the spirit of partnership achieved in coordinating this valuable element of C-SHALS with our W3C member colleagues and I hope that many of you took advantage of this excellent learning opportunity.

Finally, I thank each and every one of you for attending this small meeting. You have an opportunity to be a major contributor to the discussion topics and, therefore, to raise the value of C-SHALS for all involved. I hope you find these days to be time well spent and will plan to return next year.

Sincerely,



Burkhard Rost
Burkhard Rost
ISCB President



Welcome!

I would like to thank all the speakers for coming to present at the Second Conference on Semantics in Healthcare and Life Sciences. This conference is a unique forum for the presentation and discussion of key topics in the emerging area of semantic information technologies and their applications. We had a very successful first conference last year, and most of the participants requested we do it again this year. This conference is structured as before to be an open and exciting experience for all attendees. Our intended outcome is to engage all attendees to actively participate and contribute to the discussions. We specifically hope to identify where technologies are proving successful as well as where they still need to be developed to meet emerging scientific and business objectives. Your thought-provoking presentations will make this experience a success.

In preparation for the conference and the sessions, we asked everyone to look over the list of discussion questions (<http://www.iscb.org/cshals2009/discussion.php>) as they pertain to each participant's topical theme. These will be used to emphasize some major points as well as guide the discussions. You did not have to have written answers for all of these, but we hope you will participate by addressing those that are relevant in your area of work.

We welcome you and look forwards to meeting and speaking with each of you soon!

Sincerely,

C-SHALS Conference Committee
Eric Neumann, Clinical Semantics Group, Chair
Mike Bevil, Merck and Company
Joanne Luciano, ISCB Representative
Ted Slater, Pfizer
Susie Stephens, Eli Lilly and Company

SCHEDULE

WEDNESDAY, FEBRUARY 25

11:00 A.M.	Registration Salon IV Foyer	
12:00 NOON		
1:00 P.M.		W3C HCLSIG Tutorial Salon V – VI
2:00 P.M.		
	Break Salon V – VI Foyer	
3:00 P.M.		
4:00 P.M.	Registration Salon IV Foyer	
5:00 P.M.		Poster Reception Salon IV
6:00 P.M.		
7:00 P.M.		



SCHEDULE

THURSDAY, FEBRUARY 26

7:00 A.M.		
8:00 A.M.	Registration Salon IV Foyer	Continental Breakfast Welcome Salon IV FRAN LEWITTER <i>ISCB Board Member</i> ERIC NEUMANN <i>C-SHALS Chair</i>
9:00 A.M.		Keynote Presentation TIM BERNERS-LEE
10:00 A.M.	Break Salon IV Foyer	
11:00 A.M.		Moderated Forum 1: BioMedicine Salon IV Tim Clark Parsa Mirhaji
12:00 NOON		Tech Talk Dr. Jans Aasman Salon IV
	Lunch Salon IV Foyer	
1:00 P.M.		
2:00 P.M.		Moderated Forum 2: Semantic Web in Pharma Salon IV Dr. Philip Ashworth Ranga Chandra Gudivada Jaime Melendez Therese Vachon
3:00 P.M.	Break Salon IV Foyer	
4:00 P.M.		Moderated Forum 3: Biomarkers & Compounds Salon IV Jonas S. Almeida Thomas N. Plasterer Anthony J. Williams
5:00 P.M.		Keynote Presentation Salon IV John Reynders
	Discussion Forum and Daily Closing Remarks	
6:00 P.M.		Poster Reception Salon IV
7:00 P.M.		

SCHEDULE

FRIDAY, FEBRUARY 27

7:00 A.M.		
	Registration Salon IV Foyer	Continental Breakfast Salon IV
8:00 A.M.		
		Review – Previous Day Salon IV
9:00 A.M.		Keynote Presentation Salon IV Clark Golestani
		Discussion Forum Salon IV
10:00 A.M.	Break	Salon IV
11:00 A.M.		Moderated Forum 4: Ontologies Salon IV Marco F. Ramoni Larisa Soldatova Nigam Shah
12:00 NOON		Advanced Technologies Presentaion Salon IV Mark Wilkinson
	Lunch	Salon IV
1:00 P.M.		
		Moderated Forum 5: Knowledge Bases Salon IV Ted Slater Bruce Aronow Will Logging
2:00 P.M.		
		Closing Summary Discussion Salon IV
3:00 P.M.		
	Conference Closing Remarks & Future Actions	



KEYNOTE SPEAKERS

THURSDAY, FEBRUARY 26

9:00 A.M. – 10:00 A.M.

Linked Data in Health Care and Life Sciences: What's Next



TIM BERNERS-LEE

Director
World Wide Web
Consortium (W3C)

The Health Care and Life Sciences Community has been an early adopter of Semantic Web technologies that enable data to be linked and queried across disparate databases. While the W3C Semantic Web technology standards such as RDF, OWL, SPARQL and SKOS are being used successfully, there are still impediments to wider adoption in the health care and life sciences community. Additionally, more tools need to be developed that will enable faster and more successful use of Semantic Web technologies. We'll address what's needed and what's next from both business and technology adoption points of view.

JOHN REYNDERS

Vice-President and Chief
Information Officer
Life Sciences Division,
Johnson & Johnson

5:00 P.M. – 5:45 P.M.

Integrative Informatics: Discovery, Translational Sciences, and Epidemiology

The challenges of data management, information fusion, and knowledge mining pose a significant challenge across all phases of the pharmaceutical pipeline. As greater understanding of the composite biomarkers to delineate patient populations are elucidated, the informatics challenge involved in identifying these populations for optimal treatment has escalated. Drawing parallels with other industries, this talk will outline the scientific and technical challenges facing the pharmaceutical industry in developing the integrative informatics solutions necessary to discover and develop a next generation of personalized medicine.

KEYNOTE SPEAKERS

FRIDAY, FEBRUARY 27

8:45 A.M – 9:30 A.M.

What Must Semantics Do Now To Advance Drug Discovery?

The pharmaceutical industry finds itself at the edge of a chasm. Familiar challenges of falling research productivity, increasing costs, and tightening regulatory requirements are now being exacerbated by the global economic crisis to provide a burning platform which is driving unprecedented change. The industry is responding in a variety of ways including the formation of broader alliances with competitors, academic organizations and biotechnology companies. Remaining unchanged is the view that one cannot save one's way to prosperity, so the industry continues to make massive investments in research. These yield a vast and rapidly growing sea of data which is the industry's major asset. Yet, the challenge of deriving business value from the data is imposing a double penalty on the industry which it can ill afford. The huge expense of collecting the data continues to be compounded by another huge expense lost business opportunities which arise from inability to integrate across disparate data repositories. This situation is worsening as more broad alliances are formed fluidly across multiple organizations. Semantics hold the promise of providing the fuel for Informatics to address this key challenge, and this prospect offers a massive opportunity for the technology to impact human health. However, the need for usable enterprise semantic solutions is urgent, and the rapid delivery of the right capabilities will require a relentless focus on execution.



CLARK GOLESTANI
Vice President,
Information Technology
Merck Research
Laboratories,
Merck & Co., Inc.



HEALTH CARE AND LIFE SCIENCES INTEREST GROUP TUTORIAL

WEDNESDAY, FEBRUARY 25

11:00 a.m. – 1:00 p.m.

REGISTRATION

Salon IV Foyer

1:00 p.m. – 2:45 p.m.

W3C HCLSIG TUTORIAL

Salon V/VI

This half day tutorial will include two educational elements. W3C's Semantic Web in Health Care and Life Sciences Interest Group (HCLSIG) brings together many leading bioinformaticists and life scientists to solve a very wide array of data expression/integration problems in terms of modeling, unifying and querying biological data. This work is founded upon the full suite of W3C Semantic Web technology standards that work together to provide a rich semantic toolbox for software and system developers.

The first half of the tutorial will present a nuts-and-bolts introduction to many of the W3C Semantic Web technologies. We will motivate the introductions with use cases and examples taken from the health care and life sciences domains. The tutorial will cover the following technologies:

- RDF (data model)
- RDFS (schema language)
- OWL (ontology language)
- SPARQL (query language)
- GRDDL (XML to RDF)
- RDFa (embedding RDF in HTML)

Presenters

Lee Feigenbaum, Cambridge Semantics

HEALTH CARE AND LIFE SCIENCES INTEREST GROUP TUTORIAL



The second part of the tutorial explores how the W3C HCLSIG has been re-chartered to continue its mission to develop, advocate for, and support the use of Semantic Web technologies for biological science, translational medicine and health care. Membership in the group has grown to 89 participants, with a wide range of representation from industry and academia. The HCLSIG tutorial will discuss the challenges and opportunities at hand. An overview of the activities of the each of the current task forces in HCLSIG will be provided, along with a description of how specific Semantic Web technologies are being applied.

Presenters:

Kei Cheung, Yale University
 Tim Clark, Harvard Medical School
 Vipul Kashyap, Cigna Healthcare
 John Madden, Duke University
 Eric Prud'hommeaux, W3C
 Susie Stephens, Eli Lilly

WEDNESDAY, FEBRUARY 25

2:45 p.m. – 3:00 p.m.

BREAK

Salon V/VI Foyer

3:00 p.m. – 5:00 p.m.

W3C HCLSIG TUTORIAL

CONTINUES

Salon V/VI

4:00 p.m. – 7:00 p.m.

REGISTRATION

Salon IV Foyer

5:00 p.m. – 7:00 p.m.

POSTER RECEPTION

Salon IV



MODERATED FORUM 1

BioMedicine

THURSDAY, FEBRUARY 26

7:30 a.m. – 10:00 a.m.

REGISTRATION

Salon IV Foyer

7:30 a.m. – 8:30 a.m.

CONTINENTAL BREAKFAST

Salon IV

8:45 a.m. – 9:00 a.m.

WELCOME AND OPENING

Fran Lewitter
ISCB Board Member
Eric Neumann
C-SHALS Chair

Salon IV

9:00 a.m. – 10:00 a.m.

KEYNOTE PRESENTATION

Tim Berners-Lee

Salon IV

10:30 a.m. – 12:00 Noon

MODERATED FORUM 1

BioMedicine

Salon IV

TIMOTHY CLARK

MGH/Harvard Medical School
Boston, MA

Data, Community and Discourse: a Framework for Science Web 3.0

Science Web 3.0 is an emerging system of semantically linked data, integrated with the Social Web, in support of science collaboration and scientific knowledge integration. This system can become a powerful environment for capturing, organizing and sharing scientific knowledge, if the critical components of scientific discourse and social community are included in the framework we use to approach its design and evolution.

Such a framework must include an ontology of scientific discourse tractable to working scientists, capabilities for ontology-driven text mining, and the development of web-linked communities of scientists, computing researchers, software engineers, philanthropies and scientific publishers.

We are currently constructing Scientific Social Communities supporting three related areas of biomedical research – Alzheimer Disease, Parkinson Disease, and stem cell research – using such a framework, through an expanding series of collaborations. We will discuss these projects, and present our current software, ontology and activity-theoretic models for the development of this work.



MODERATED FORUM 1

BioMedicine

PARSA MIRHAJI

The University of Texas Health Science Center
Houston, Texas

Biomedical Language Understanding and Extraction (BLUE-Text): A Minimal Syntactic, Semantic Method

Although many techniques have been introduced for processing of unconstrained text in clinical settings, natural language processing (NLP) is not yet conceived as an integral component of electronic health record systems, and its utilization and adoption does not match its potentials. Furthermore, the translational research initiatives call for generation of technologies that integrate clinical text with structured data not only for queries, and information retrieval but also for contextualization, and reuse of clinical information for multidisciplinary research in a collaborative and distributed environment envisioned by CTSA program.

The Biomedical Language Understanding and Extraction system (BLUE-Text) is a minimal syntactic, semantic algorithm that aims at bridging some of the gaps between the current and desired state of the art for the processing of unconstrained clinical text, and aims at constructing a dynamically flexible, customizable (extensible), consistent, formal, and explicit representation of clinical text suitable for information sharing and reuse.

**THURSDAY,
FEBRUARY 26**

10:30 a.m. – 12:00 Noon

**MODERATED FORUM 1
BioMedicine**

Salon IV



MODERATED FORUM 2

Semantic Web in Pharma

THURSDAY, FEBRUARY 26

12 noon – 12:10 p.m.

TECH TALK Dr. Jans Aasman CEO, Franz Inc.

Salon IV

12:10 pm – 1:30 p.m.

LUNCH

Salon IV

1:30 p.m. – 3:00 p.m.

MODERATED FORUM 2 Semantic Web in Pharma

Salon IV

PHILIP ASHWORTH

UCB

Berkshire, UK

Cool URIs for Molecules

Semantic web standards are designed with the data integration in mind. Key integration capabilities foundational to RDF include globally unique identifiers and the ability to merge information. Data integration is of high importance in life sciences where a wealth of different datasets contain separate, but related information.

Integration can produce important new insights but existing technologies struggle to achieve this goal. As a result, interest in semantic technologies is high. While technology standards are important in supporting integration goals, so is the adoption of industry standards. This presentation will investigate and propose potential Semantic Web standards that could be applied to chemicals entities. We will highlight how this approach has been used to integrate chemical datasets without the need for complex chemically intelligent software packages.

In this talk we will:

- Describe the nature of the problem
- Propose a standard for chemical entities
- Highlight tools used
- Demonstrate integration of chemical entities



MODERATED FORUM 2

Semantic Web in Pharma

RANGA CHANDRA GUDIVADA

Eli Lilly and Company
Indianapolis, IN

Building a Pharmaceutical Competitive Landscape using Semantic Technologies

Building a 'Pharmaceutical Competitive Landscape' (CL) is a purposeful and co-ordinated monitoring of the competitors in the industry within a specific market place. Strategically, this is to gain foreknowledge and stay abreast of recent developments of your competitor's plans in order to make informed decisions and to plan your business strategy to countervail their plans. It is a systematic ethical program that involves information aggregation, integration into your existing knowledge infrastructure.

This process helps to make a calculated business plans, decisions and operations on the grounds of high-quality information, analysis and formulation of hypothesis. The raw data is usually available as vendor feeds and is syntactically and semantically varied. Knowledge discovery from these data sets is possible only through extensive semantic integration and ability to pose complex queries. Here, we present a case study in endocrine area, where we applied a semantic structure to raw text from different vendors by mapping to domain specific ontologies using simple heuristics. In addition, we adopted Semantic Web standards for knowledge representation and querying due to its support for rich semantics and inference capabilities.

**THURSDAY,
FEBRUARY 26**

1:30 p.m. – 3:00 p.m.

MODERATED FORUM 2

**Semantic Web in
Pharma**

Salon IV

**MODERATED FORUM 2****Semantic Web in Pharma****THURSDAY,
FEBRUARY 26**

1:30 p.m. – 3:00 p.m.

**MODERATED FORUM 2
Semantic Web in
Pharma**

Salon IV

JAIME MELENDEZMerck and Co, Inc.
Westpoint, PA***Semantic Web Technologies Within
Research and Development***

The utilization of semantic technologies and agile development has enabled the creation of a domain information model and provide a facile mechanism to combine data from multiple sources into a single integrated knowledge base — critical in providing insight into the quality and targets associated with the collection of assay information. By providing this integrated knowledge base, concept based search and visual navigation are applied seamlessly integrating data stores resulting in an improved search precision and expressiveness. The capability to gain understanding of the structure and function based on this information is essential in order to effectively develop novel products of quality.

**MODERATED FORUM 2****Semantic Web in Pharma****THERESE VACHON**

Novartis Pharma AG
Basel, Switzerland

***Bridging Knowledge Gaps in Drug
Discovery with Semantic Technologies***

In the Life Sciences, there is a growing need to provide tools for answering complex queries to support the drug discovery process. Hundreds of different questions have already been identified in the course of a data federation project in NIBR (Novartis Institutes for Biomedical Research). Technologies that support the semantic layer will be presented as well as various tools for accessing, aggregating and analyzing disparate sets of data.

**THURSDAY,
FEBRUARY 26**

1:30 p.m. – 3:00 p.m.

**MODERATED FORUM 2
Semantic Web in
Pharma**

Salon IV



MODERATED FORUM 3

Biomarkers & Compounds

**THURSDAY,
FEBRUARY 26**

3:00 p.m. – 3:30 p.m.

BREAK

Salon IV

3:30 p.m. – 5:00 p.m.

MODERATED FORUM 3

**Biomarkers &
Compounds**

Salon IV

JONAS ALMEIDA

The University of Texas M. D. Anderson Cancer Center
Houston, TX

Cancer Biomarker Research — Integrating Sensitive, Heterogeneous, and Labile, Experimental Data Sources with Semantic Core Models that Embed User Permissions

The maturation of semantic web technologies (SW) offers a more generic foundation to weave integrated data management systems than the relational and object oriented approaches that precede it. Because the meaning and representation of molecular biology data changes with the results of its analysis, data management and data analysis cannot be treated as separate components of knowledge engineering for the Life Sciences. A distributed triple store application was developed to address this issue. The individual nodes of this distributed infrastructure, designated as S3DB (available at S3DB.org), include a RDFS core model with a permission scheme embedded. The interoperation between nodes is achieved through REST web services supporting SPARQL. The resulting infrastructure was evaluated for the merging of data management systems across two cancer centers of the University of Texas: MDAnderson Cancer Center at Houston and the Southwestern Medical Center at Dallas. Particular attention was put on issues of security and privacy, which are central to biomarker discovery initiatives. The more generic nature of SW abstractions enabled the incorporation of access permission in the data model such that permission to access data elements can travel with the data rather than staying with the point of access to the individual data store.

MODERATED FORUM 3

Biomarkers & Compounds

THOMAS PLASTERER

BG Medicine
Waltham, MA

Enabling Discovery in High-Risk Plaque using Semantic Web Approaches

The HRP initiative (HRP) is a joint research and development effort to advance the understanding, recognition and management of high-risk plaque for the benefit of multiple stakeholders in the healthcare system. As the primary underlying cause of heart attacks, high-risk, or vulnerable plaque is the number one cause of death in the Western world. There are currently no methods of screening, diagnosis or treatment for high-risk plaque.

The HRP initiative leverages recent advances in biology and information technology to design and optimize a care-cycle for high-risk plaque, promising to reduce morbidity, mortality and cost associated with cardiovascular disease. This Initiative is being led by the world's foremost scientists in the fields of cardiology, pathology, and imaging, and is made possible through funding by leading pharmaceutical and medical technology entities.

HRP takes advantages of semantic web technologies for physician and researcher-lead data analysis and data interoperability. One of the key applications is a web tool linking patient demographics, clinical chemistries, physical measurements and cardiovascular imaging modalities. This empowers scientists to rapidly compare multiple clinical parameters to find patients of interest, assisting greatly in defining high-risk plaque.

THURSDAY, FEBRUARY 26

3:30 p.m. – 5:00 p.m.

MODERATED FORUM 3 Biomarkers & Compounds

Salon IV



MODERATED FORUM 3

Biomarkers & Compounds

THURSDAY, FEBRUARY 26

3:30 p.m. – 5:00 p.m.

MODERATED FORUM 3 Biomarkers & Compounds

Salon IV

5:00 p.m. – 5:45 p.m.

KEYNOTE PRESENTATION John Reynders

Salon IV

5:45 p.m. – 6:00 p.m.

DISCUSSION FORUM AND DAILY CLOSING REMARKS

Salon IV

6:00 p.m. – 7:30 p.m.

POSTER RECEPTION

Salon IV

ANTONY WILLIAMS

ChemZoo Inc.
Wake Forest, North Carolina

Crowdsourcing, Collaborations and Text-Mining in a World of Open Chemistry

The increasing availability of free and open access resources for scientists on the internet presents us with a revolution in data availability. However, freedom costs and in many cases the cost is quality. ChemSpider is a free access website for chemists built with the intention of providing a structure centric community for chemists. As an aggregator of chemistry related information from many sources, at present over 21.5 million unique chemical entities from over 150 separate data sources, ChemSpider has taken on the task of both robotically and manually curating publicly available data sources. This presentation will provide an overview of how a curated platform can become the centralized hub for resourcing information about chemical entities. We will also present ChemMantis, an entity extraction platform for extracting chemical names and scientific terms in documents and providing a platform for structure-based searching of Open Access chemistry literature.

**MODERATED FORUM 4****Ontologies****MARCO RAMONI**

Harvard Medical School
Cambridge, MA

Automated Ontology Engineering

Every year, over 400,000 new articles enter the biomedical literature. This staggering growth has motivated the development of ontologies, structured information repositories organizing biomedical findings into hierarchical structures. These ontologies have become even more important with the introduction of high-throughput genomic platforms because of their ability to categorize large amounts of information using a controlled vocabulary. Today, ontologies are developed through manual curation, a highly laborious, costly and slow process of consensus creation among domain experts. With the increased complexity of biomedical data, the scalability of ontology development will require automated methods to support and streamline the curation process. This talk will describe the principles and the challenges of developing methods for the automated engineering of ontologies and, using the evolution of Gene Ontology, the most commonly used and influential biomedical ontology, it will show that the principled engineering of ontologies leads to more accurate analytical results.

**FRIDAY,
FEBRUARY 27**

7:30 a.m. – 10:00 a.m.

REGISTRATION

Salon IV Foyer

7:30 a.m. – 8:30 a.m.

**CONTINENTAL
BREAKFAST**

Salon IV

8:30 a.m. – 8:45 a.m.

**REVIEW –
PREVIOUS DAY**

Salon IV

8:45 a.m. – 9:30 a.m.

KEYNOTE PRESENTATION**Clark Golestani**

Salon IV

9:30 a.m. – 10:00 a.m.

DISCUSSION FORUM

Salon IV

10:00 a.m. – 10:30 a.m.

BREAK

Salon IV

10:30 a.m. – 12:00 noon

MODERATED FORUM 4**Ontologies**

Salon IV



MODERATED FORUM 4

Ontologies

**FRIDAY,
FEBRUARY 27**

10:30 a.m. – 12:00 noon

**MODERATED FORUM 4
Ontologies**

Salon IV

LARISA SOLDATOVA

Aberystwyth University
Aberystwyth, UK

The Formal Description of Drug Screening and Design Investigations

The Computational Biology group at the University of Wales, Aberystwyth has been developing technologies for the automation of scientific discovery for more than a decade. We have introduced a concept of a Robot Scientist, a physically implemented system to run cycles of automated scientific investigations. These investigations involve cycles of: hypotheses formation, design of experiments, physically execution of the experiments, analysis of the results, and updating of background knowledge. The automation of investigations requires a detailed and fully formalized machine-readable representation. Robot Scientist investigations are represented and recorded using an ontology LABORS (LABoratory Ontology for Robot Scientists) which defines all essential components of the investigations.

We are now applying the Robot Scientist approach to the automation of drug screening and design. The Robot Scientist “Eve” has been designed to automatically run investigations that target 3rd world diseases using yeast strains that express chimeric GPCRs. The design features which distinguish Eve from existing high-throughput screening automation are: the ability of Eve during the screening process to decide to switch to QSAR mode; integration of machine learning and cycles of active learning to learn QSARs; and a relational approach for the representation of compound structure/substructure. Currently Eve is in the final stage of being built and SATs are scheduled for December.

MODERATED FORUM 4

Ontologies

Eve has a processing station where a range of high-throughput assays are prepared. The station has incubators to hold yeast strains, storage facilities for the compound library (~15,000 compounds), plate transport robotics to move plates to liquid handling, assay stations, a liquid handling station to enable high-throughput assay preparation, a set of fluorescent plate readers, integration hardware, an automated microscope, computers, and associated software.

In order to support automated drug screening and design investigations we are developing the ontology DDI (Drug screening and Design Investigations) based on LABORS. DDI defines the structure of relevant investigations, objects of investigation, targets, observations, results, what data and metadata to record, equipment and its functionality, materials, etc. DDI is used to design a data base to store the defined data and metadata from experiments. Both the ontology and the database can be translated into Datalog in order to enable computational inference over all the collected data, metadata, and their underlying structure defined in DDI. DDI is also used for the formalized representation of experiment protocols: screening, 'cherry picking', and QSAR. Each experimental action, its properties, pre- and post-conditions are fully defined.

Currently DDI is sufficient for the description of our drug screening and design investigations. The design of DDI is compliant with OBI (the Ontology for Biomedical Investigations). We propose to extend DDI to the whole drug discovery pipeline. We argue that an extended ontology will be useful for the developing standards to record/store/retrieve/exchange information about drug design. We therefore invite interested researchers and practitioners to participate in the DDI development and extension.

FRIDAY, FEBRUARY 27

10:30 a.m. – 12:00 noon

MODERATED FORUM 4 Ontologies

Salon IV



MODERATED FORUM 4

Ontologies

**FRIDAY,
FEBRUARY 27**

10:30 a.m. – 12:00 noon

MODERATED FORUM 4
Ontologies

Salon IV

NIGAM SHAH
Stanford University
Stanford, CA

Ontology Services for Semantic Application in Health and Life Sciences

The National Center for Biomedical Ontology, one of the seven National Centers for Biomedical Computing (NCBC) created under the NIH Roadmap, is developing BioPortal, a Web-based system that serves as a repository for biomedical ontologies. BioPortal provides access to one of the largest libraries of biomedical ontologies both via Web browsers and Web services. BioPortal also offers community-based evolution of ontology content, which has supported the evolution of the Biomedical Resource Ontology developed jointly by the NCBCs. BioPortal enables ontology users to learn what biomedical ontologies exist, what a particular ontology might be good for, and how individual ontologies relate to one another. The BioPortal Web services allow users to access, download, traverse, and map ontologies to one another. The Web services also allow users to submit textual metadata from their databases and automatically “tag” the metadata with ontology terms. Our Web services have been used to create a searchable ontology-based index of elements from PubMed, ClinicalTrials.gov, and the Gene Expression Omnibus.



MODERATED FORUM 5

Knowledge Bases

TED SLATER

Pfizer Inc.
Chesterfield, MO

Practical Knowledge Engineering for Pharmaceutical R&D

Successful pharmaceutical R&D requires the effective use of all our knowledge. Most informatics efforts are aimed at data integration, which is difficult or impossible to achieve, but what is seldom appreciated is the fact that data integration is simply not enough to enhance our ability to accurately predict drug efficacy and safety. To reach the next level in pharmaceutical R&D, we must move beyond data integration to the point where our computers generate hypotheses for us. This talk will provide attendees with a view of a practical, semantics-based solution which provides for automated reasoning over pharmaceutical knowledge bases, and which has useful emergent properties such as streamlined data interoperability and a universal platform for knowledge management.

FRIDAY, FEBRUARY 27

10:30 a.m. – 12:00 noon

MODERATED FORUM 4 Ontologies

Salon IV

12:00 noon – 12:25 p.m.

ADVANCED TECH TALK Mark Wilkinson

Salon IV

12:25 p.m. – 1:30 p.m.

LUNCH

Salon IV

1:30 p.m. – 3:00 p.m.

MODERATED FORUM 5 Knowledge Bases

Salon IV



MODERATED FORUM 5

Knowledge Bases

**FRIDAY,
FEBRUARY 27**

1:30 p.m. – 3:00 p.m.

MODERATED FORUM 5
Knowledge Bases

Salon IV

BRUCE ARONOW

Cincinnati Children's Hospital Medical Center
University of Cincinnati
Cincinnati, Ohio

Applied Diseaseomics: Can Assembly and Analysis of Disease-related Feature Networks Enable Inferential Reasoning for Causal Factors and Candidate Therapeutics?

This talk will outline methods for rapid network-based aggregation of current knowledge associated with specific diseases or disease types. By extending connections of known associated genes, phenotypes, or similar diseases, a variety of comparative feature enrichment analyses can be performed that allow connectivity to pathways, etiologies, and therapeutics. Using domain expertise, reasoning-based approaches are clearly enabled that can shed light into operative or inoperative biological processes that underlie disease processes. These analytical and modeling approaches can then help to suggest candidate modifiers that may be systematically evaluated such as variant genetics, etiologic factors, and novel therapeutic agents. An important question is how much of this data integration, analysis, interpretation, and hypothesis process could be empowered using semantic web-based methodologies?



MODERATED FORUM 5

Knowledge Bases

WILLIAM LOGING

Boehringer Ingelheim Pharmaceuticals, Inc.
Ridgefield, CT

High-Throughput Electronic Biology

Moore's Law documents how increases in computational platform power increase exponentially. The applications of this rule for computer science to fields such as physics and Artificial Intelligence are evident; given the ever evolving nature of computing technology. The current in silico "data deluge" that has taken place within the chemical/genomics revolution presents the computational research field with an unparalleled opportunity — to move beyond simple data integration to the emergent field of Electronic Biology; i.e. the developing field of science in which computer science and biology merge into a single discipline. As larger and more complex datasets become available within the Electronic Biology field, newer methods of capturing and analyzing data will be required for discovery.

FRIDAY, FEBRUARY 27

1:30 p.m. – 3:00 p.m.

MODERATED FORUM 5 Knowledge Bases

Salon IV

3:00 p.m. – 3:30 p.m.

CLOSING SUMMARY DISCUSSION

Salon IV

3:30 p.m. – 3:45 p.m.

CONFERENCE CLOSING AND FUTURE ACTION

Salon IV



TECH TALKS

THURSDAY, FEBRUARY 26

12:00 noon – 12:10 p.m.

TECH TALK

Salon IV

FRIDAY, FEBRUARY 27

12:00 noon – 12:25 p.m.

ADVANCED TECH TALK

Salon IV

Tech Talk

DR. JANS AASMAN

CEO, Franz Inc.

Optimizing SPARQL for Diverse Ontologies

Advanced Technologies Presentation

DR. MARK WILKINSON

IO Informatics

CardioSHARE

The Semantic Automated Discovery and Integration (SADI) framework transparently exposes Web Service output as a Semantic Web resource Application towards medical advances in cardiac diseases. (<http://cardioshare.icapture.ubc.ca/>)

SADI is a technology designed to expose Web Services and their output as queryable Semantic Web resources. We demonstrate SADI as a plug-in to the IO Informatics Knowledge Explorer user-interface as an example of generating disease-specific knowledgebases to advance understanding of heart disease.

ABSTRACT: The “Deep Web” — all of the Web data accessible only through Web forms — is believed to contain hundreds to thousands of times more data than the Web itself. Here we demonstrate that Deep Web data, when exposed as Semantic Web Services, can be discovered, explored, and queried as if it were a typical Semantic Web resource. Our SADI framework interprets SPARQL queries, determines which Semantic Web Services are capable of generating the data needed to answer that query, and then



TECH TALKS

executes those services in order to dynamically generate the database used for query resolution. Moreover, well-defined OWL-DL classes referred to in the query are deconstructed to determine their defining property-restrictions, and these properties are also used for Semantic Web Service discovery. This allows “naive” data to be analyzed for potential classification into these OWL classes, and thus utilized for query answering. We have created an implementation of SADI as a plug-in to the Knowledge Explorer interface from IO Informatics. This provides an intuitive way for end-users to visualize and explore the data retrieved from SADI, as well as execute SADI Semantic Web Services through a simple query interface. The Use and application of this approach towards medical advances in the CardioSHARE (<http://cardioshare.icapture.ubc.ca/>) initiative are demonstrated and discussed. This presentation demonstrates the utility and power gained by utilizing the property restrictions in OWL-DL, rather than creating simple asserted class-hierarchies. Moreover, it demonstrates the ability of machines on the Semantic Web to re-interpret and utilize data in ways that were not anticipated by the data provider.

FRIDAY, FEBRUARY 27

12:00 noon – 12:25 p.m.

ADVANCED TECH TALK

Salon IV



POSTERS

WEDNESDAY, FEBRUARY 25

5:00 p.m. – 7:00 p.m.

POSTER RECEPTION

Salon IV

THURSDAY, FEBRUARY 26

6:00 p.m. – 7:30 p.m.

POSTER RECEPTION

Salon IV

Structural Database Using Semantic Web Concepts to Support Structure-Based Drug Design for AIDS

T. N. BHAT

NIST

bhat@nist.gov

The HIV structural databases (HIVSDB, http://bioinfo.nist.gov/SemanticWeb_pr2d/chemblast.do, <http://chemdb2.niaid.nih.gov>) distribute one of the largest comprehensive collections of structural, biological and pre-clinical data on inhibitors, drug leads and clinical drugs for AIDS. These databases contain info on several thousand biologically active compounds from all classes (HIV PR, RT, CCR5, Integrase) of FDA approved drugs. Efficient and yet user friendly data management systems that support state-of-the-art annotation, visualization and query capabilities are crucial for the effective use of data for fragment based structural pharmacology and rational drug design. Semantic Web is the vision of the World Wide Web Consortium for enabling seamless integration of electronic data for data mining and knowledge generation across the Web.

Robust and functionally relevant ontology plays a critical role in developing the data elements for a Semantic Web. Presentation will illustrate how Semantic Web concepts are used for novel annotation, data integration, storage, and query to manage and display structural (fragments, 2-D images and text-based) biological, and pre-clinical data. One of these techniques (ChemBLAST (Prasanna, Vondrasek et al. 2006)) developed allows rapid comparison of structural fragments using the Semantics commonly used in drug discovery process. At present majority of the data in HIVSDB are obtained by us by weaning through publications. Our intension is to seek greater participation by the community by depositing data to HIVSDB at the time of publication.

Prasanna, M. D., J. Vondrasek, et al. (2006). "Chemical compound navigator: a web-based chem-BLAST, chemical taxonomy-based search engine for browsing compounds." *Proteins* 63(4): 907-17.

POSTERS

Exposing The Cancer Genome Atlas (TCGA) as a SPARQL Endpoint

HELENA F DEUS

The University of Texas M. D. Anderson Cancer Center
mhdeus@mdanderson.org

Automated discovery of candidate biomarkers from multiple databases has been a central challenge in the Life Sciences in general and in the study of systemic processes such as cancer biology in particular. The maturation of Semantic Web technologies offers solutions to those problems by allowing the query to be defined by the domains of discourse where the answer to the query is sought. A specific example of this challenge is found in The Cancer Genome Atlas initiative (TCGA, <http://cancergenome.nih.gov/>), which generates a large scale repository of high throughput molecular biology data generated and processed at 5 academic facilities across the USA [1, 2]. The heterogeneity of domains (genomic, transcriptomic, epigenetic effects, proteomic, clinic and demographic, etc) and the heterogeneity of methodologies within each domain will be further compounded by the expansion of TCGA as an international initiative lead by The International Genomics Consortium (IGC) in collaboration with the Translational Genomics Research Institute (TGen). Currently, the TCGA data is aggregated at a single point of access charged with providing syntactic interoperability to all of the data the TCGA portal.

Using The Cancer Genome Atlas as a case study, and the S3DB (www.s3db.org, [3, 4]) distributed semantic data service application as the engine of integration, we developed a computational domain representation for the TCGA data in order to integrate the clinical and molecular information and expose it through Web Services. Specifically, this novel resource allows information retrieval through the SPARQL module available at any S3DB node deployment. Since sensitive and proprietary data is always a sensitive preoccupation with translational studies in Biomedicine, the ontology itself includes the management of user permissions on individual data elements. The architecture for this novel resource is thought to provide a template for web-based solutions that bridge between data

WEDNESDAY, FEBRUARY 25

5:00 p.m. – 7:00 p.m.

POSTER RECEPTION
Salon IV

THURSDAY, FEBRUARY 26

6:00 p.m. – 7:30 p.m.

POSTER RECEPTION
Salon IV



POSTERS

WEDNESDAY, FEBRUARY 25

5:00 p.m. – 7:00 p.m.

POSTER RECEPTION

Salon IV

THURSDAY, FEBRUARY 26

6:00 p.m. – 7:30 p.m.

POSTER RECEPTION

Salon IV

silos within a domain of knowledge and between the bench and the clinical point of care.

[1] Cancer Genome Atlas Research Network. "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, pp. 1061-8, Oct 23 2008.

[2] L. Chin et al. "Translating insights from the cancer genome into clinical practice," *Nature*, vol. 452, pp. 553-63, Apr 3 2008.

[3] J. S. Almeida et al. "Data integration gets sloppy," *Nat Biotechnol*, vol. 24, pp. 1070-1, Sep 2006.

[4] H. F. Deus et al. "A Semantic Web Management Model for Integrative Biomedical Informatics," *PLoS ONE*, vol. 3, p. e2946, 2008.

SPONSOR HIGHLIGHTS

IO Informatics is at the forefront of a global revolution in software methods for data integration. IO Informatics' Sentient suite of software products enables biotechnology, pharmaceutical, medical and other life science researchers to structure and define complex data relationships, view these relationships, query them, and capitalize on them — all within a secure, compliant, auditable framework that helps organizations accumulate and leverage knowledge. Sentient software brings together a unique combination of Semantic and search technologies coupled with data and process management to improve efficiency and deliver more meaningful results to Life Science and Healthcare research

organizations. Founded in 2003, IO Informatics is headquartered in Berkeley, California (www.io-informatics.com).

IO Informatics is pleased to be a sponsor of this ground breaking and important conference! www.io-informatics.com

(+1) 510-705-847



POSTERS

BioProspecting the Bibliome: Discovering Potentially Novel Cancer Biomarkers

PETER ELKIN

Mount Sinai Medical Center

peter.elkin@mssm.edu

Using SNOMED-CT and HUGO as ontologies, 27,000 Web-accessible NEJM articles were mined for potentially novel biomarkers using NLP. Articles containing an association between a gene and a metabolic function were linked with articles associating the same metabolic function and a disease, and no single article associated the same gene and disease (i.e. novel association).

Building Innovation Networks: Applying Semantic Technology in the Life Sciences

MARC HADFIELD

Alitora Systems, Inc.

marc@alitora.com

To be successful, Life Science professionals must utilize the perfect storm of heterogeneous data – genomic research, clinical studies, health care records, patents, industry news, market research, and government policy.

Semantic Technology provides a means of linking related heterogeneous data assets together using Information Extraction and other Data Mining techniques aligned using Ontologies. People too are an important part of the Innovation Network – much of an organization's data is locked in the brains of their experts.

In this presentation, we will explore specific tools and techniques to build Innovation Networks using the Memomics application as a Case Study. A demonstration of the Memomics Innovation Network will combine data assets from heterogeneous sources with expert users. Users of the Innovation Network will be able to collaborate with other users, sharing knowledge and expertise.

The Memomics Innovation Network uses Semantic Technology to expose data assets to semantic searches,

WEDNESDAY, FEBRUARY 25

5:00 p.m. – 7:00 p.m.

POSTER RECEPTION

Salon IV

THURSDAY, FEBRUARY 26

6:00 p.m. – 7:30 p.m.

POSTER RECEPTION

Salon IV



POSTERS

WEDNESDAY, FEBRUARY 25

5:00 p.m. – 7:00 p.m.

POSTER RECEPTION

Salon IV

THURSDAY, FEBRUARY 26

6:00 p.m. – 7:30 p.m.

POSTER RECEPTION

Salon IV

and associates discovered knowledge with the users that may be most interested in it.

Issues will be explored such as:

- Accuracy of Semantic data, and Errors
- Semantic Searches
- Matching users to knowledge
- Utilizing Multiple Ontologies
- Collaboration across Organizations
- Security

Semantic Integration in Biomedical Networks

CHUN-HSI HUANG

University of Connecticut
huang@cse.uconn.edu

A semantic network is a conceptual model for knowledge representation, in which the knowledge entities are represented by nodes (or vertices), while the edges (or arcs) are the relations between entities. The semantic network is an effective tool, serving as the backbone knowledge representation system for genomic, clinical and medical data. Usually these knowledge bases are stored at locations geographically distributed. This highlights the importance of an efficient distributed semantic network system enabling distributed knowledge extraction and inference. Note that the semantic network is a key component of the Unified Medical Language System (UMLS) project initiated in 1986 by the U.S. National Library of Medicine (NLM). The goal of the UMLS is to facilitate associative retrieval and integration of biomedical information so researchers and health professionals can use such information from different (readable) sources. The UMLS project consists of three core components: (1) the Metathesaurus, providing a common structure for more than 95 source biomedical vocabularies. It is organized by concept, which is a cluster of terms, e.g., synonyms, lexical variants, and translations, with the same meaning; (2) the Semantic Network, categorizing these concepts by semantic types and relationships; and (3) the SPECIALIST lexicon and associated lexical tools, containing over 30,000 English words, including various biomedical terminologies. Information for each entry, including base form, spelling variants, syntactic category,



POSTERS

inflectional variation of nouns and conjugation of verbs, is used by the lexical tools. The 2002 version of the Metathesaurus contains 871,584 concepts named by 2.1 million terms. It also includes inter-concept relationships across multiple vocabularies, concept categorization, and information on concept co-occurrence in MEDLINE.

Application areas in biomedicine include the epidemiological studies and medical imaging, which produce tremendous amount of data that are usually geographically distributed among hospitals, clinics, research labs, and radiology centers, etc. For research, training or clinical purposes, physicians and researchers often need to consult and analyze data from distributed sites. Thus, an infrastructure supporting on-demand and automated information extraction and reasoning will provide significant convenience.

This research work involves a few tasks, including (1) the development of a distributed semantic network system, based on a task-based and message-driven model to exploit both task and data parallelism while processing queries; (2) the parallelization of the

WEDNESDAY, FEBRUARY 25

5:00 p.m. – 7:00 p.m.

POSTER RECEPTION
Salon IV

THURSDAY, FEBRUARY 26

6:00 p.m. – 7:30 p.m.

POSTER RECEPTION
Salon IV

SPONSOR HIGHLIGHTS

Founded in 1849, Pfizer is the world's largest research-based pharmaceutical company taking new approaches to better health. We discover, develop, manufacture and deliver quality, safe and effective prescription medicines to treat and help prevent disease for both people and animals. We also partner with health care providers, governments and local communities around the world to expand access to our medicines and to provide better quality health care and health system support. At Pfizer, nearly 90,000 colleagues in more than 150 countries work every day to help people stay happier and healthier longer and to reduce the human and economic burden of disease worldwide.





POSTERS

WEDNESDAY, FEBRUARY 25

5:00 p.m. – 7:00 p.m.

POSTER RECEPTION

Salon IV

THURSDAY, FEBRUARY 26

6:00 p.m. – 7:30 p.m.

POSTER RECEPTION

Salon IV

inference engine to speed-up the query processing; and (3) automated data migration among the distributed knowledge bases to maximize the storage utilization rate. The current information infrastructure, as a test-bed, of this project is a campus-wide computational and data Grid. Participating sites of this infrastructure include the Schools of Engineering and Medicine at the University of Connecticut. Note that the Grid represents a rapidly emerging and expanding technology that allows geographically distributed resources (CPU cycles, data storage, sensors, visualization devices, and a wide variety of internet-ready instruments), which are under distinct control, to be linked together in a transparent fashion. The aggregate computing power, data storage, network bandwidth, as well as the user friendliness have rendered the Grid a prosperous infrastructure in support of automated processing of distributed information.

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





THANK YOU TO OUR SPONSORS

GOLD

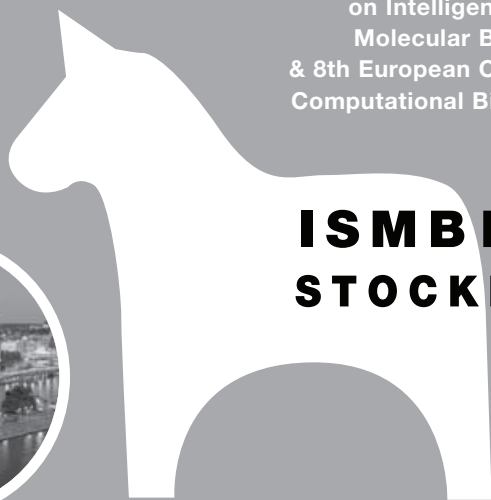


BRONZE



An official conference of the
International Society for
Computational Biology

17th Annual International Conference
on Intelligent Systems for
Molecular Biology (ISMB)
& 8th European Conference on
Computational Biology (ECCB)



ISMB/ECCB STOCKHOLM 2009

JUNE 27 – JULY 2

CONFERENCE CHAIRS

Gunnar von Heijne,

Honorary Conference
Chair, *Stockholm
University, Stockholm,
Sweden*

Eugene Myers,
Conference Co-chair
*HHMI Janelia Farm
Research Campus,
Ashburn, VA, USA*

Marie-France Sagot,
Conference Co-chair
*INRIA Grenoble —
Rhône-Alpes Research
Centre, Montbonnot-
Saint Martin, France*

Pierre-Henri Gouyon,
*Evolution et Systématique des Végétaux,
Université de Paris-Sud, Orsay, France*

Daphne Koller,
Department of Computer Science,
Stanford University, Stanford, USA

Thomas Lengauer,
Max-Planck Institute for Informatics,
Saarbrücken, Germany

Tomaso A. Poggio,
Eugene McDermott Professor, Brain and
Cognitive Sciences, McGovern Institute
for Brain Research, Computer Science
and Artificial Intelligence Laboratory,
*Massachusetts Institute of Technology,
Cambridge, USA*

Mathias Uhlen, *Royal Institute of
Technology (KTH), Stockholm, Sweden*

**Eugenia María del
Pino Veintimilla,**
*Pontifical Catholic University of Ecuador
(PUCE), Quito, Ecuador*

ISCB OVERTON PRIZE
Trey Ideker, Department of
Bioengineering, *UC San Diego, USA*

2009 ISCB SENIOR SCIENTIST
ACCOMPLISHMENT AWARD
Webb Miller,
*Pennsylvania State University,
University Park, USA*

Register
Now!



<http://www.iscb.org/ismbecb2009/>





An Official Conference
of the International Society
for Computational Biology

www.iscb.org