Conference on Semantics in Healthcare and Life Sciences

February 22–24, 2012
ROYAL SONESTA HOTEL BOSTON
CAMBRIDGE/BOSTON, USA
ISCB gratefully acknowledges our sponsors for their support and commitment to this conference.

**GOLD**

**MERCK**

**GENERAL**

**O’REILLY®**

**TWC**

**Rensselaer**

**STUDENT TRAVEL FELLOWSHIP**

**AstraZeneca**

An official conference of the International Society for Computational Biology
Welcome to CSHALS!

Thank you for choosing to attend the 5th Annual ISCB Conference on Semantics in Healthcare and Life Sciences. By the end of this event I hope you have a better understanding of Big Data challenges and solutions, including specific applications of cutting edge semantic technologies for the pharmaceutical and healthcare industries. The conference format is specifically designed to encourage you to engage our visionary speakers and one another so you may take home new ideas, new methods, and new contacts to draw from in the future.

For those of you who might be new to using intelligent information technologies in pharmaceutical R&D, I hope you took advantage of the pre-conference tutorial organized by the Rensselaer Polytechnic Institute’s Tetherless World Constellation (RPI). Like last year, ISCB is grateful to RPI for customizing a hands-on tutorial to help CSHALS attendees come quickly up to speed on this cutting edge technology.

This year’s conference was organized under the leadership of conference co-chairs Mike Bevil (Merck) and Joanne Luciano (RPI), together with organizing committee members Jonas Almeida, Lee Feigenbaum, and Ted Slater. ISCB’s Director of Conferences, Steven Leard, managed all logistics while ISCB’s Executive Officer, BJ Morrison McKay, played a supporting role in ensuring 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 you for attending, and hope that CSHALS 2012 exceeds your expectations.

Enjoy!

Burkhard Rost
ISCB President
Welcome!

On behalf of the entire CSHALS 2012 Organizing Committee, we would like to thank all the speakers for agreeing to present at the Fifth 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,

Mike Bevil,
Merck

Joanne S. Luciano,
Rensselaer Polytechnic Institute

CSHALS Conference Co-chairs
Wednesday, February 22

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:30 AM – 1:00 PM</td>
<td>Registration</td>
</tr>
<tr>
<td>8:30 AM – 10:30 AM</td>
<td>RPI-LED HANDS-ON TUTORIAL • Semantic Healthcare and Life Sciences Tutorial: Mashing HC and LS Data</td>
</tr>
<tr>
<td></td>
<td>CSHALS Tutorial Coordinator: Lee Feigenbaum, Cambridge Semantics</td>
</tr>
<tr>
<td></td>
<td>RPI Tutorial Coordinator: Joanne S. Luciano</td>
</tr>
<tr>
<td></td>
<td>Presenters: Dominic DiFranzo, James McCusker, Joshua Shinavier</td>
</tr>
<tr>
<td>10:30 AM – 11:00 AM</td>
<td>Break</td>
</tr>
<tr>
<td>11:00 AM – 12:15 PM</td>
<td>RPI-LED HANDS-ON TUTORIAL (continues)</td>
</tr>
<tr>
<td>12:15 PM – 1:00 PM</td>
<td>Lunch</td>
</tr>
<tr>
<td>1:00 PM – 3:00 PM</td>
<td>RPI-LED HANDS-ON TUTORIAL (continues)</td>
</tr>
<tr>
<td>3:00 PM – 3:15 PM</td>
<td>Break</td>
</tr>
<tr>
<td>3:15 PM – 5:00 PM</td>
<td>RPI-LED HANDS-ON TUTORIAL (continues)</td>
</tr>
<tr>
<td>4:00 PM – 7:00 PM</td>
<td>Registration</td>
</tr>
<tr>
<td>4:00 PM – 5:00 PM</td>
<td>Poster (Author) Set-up</td>
</tr>
<tr>
<td>5:00 PM – 7:00 PM</td>
<td>POSTER RECEPTION</td>
</tr>
</tbody>
</table>

SPONSOR SPOTLIGHT

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.
### Thursday, February 23

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:30 AM – 10:00 AM</td>
<td>Registration</td>
</tr>
<tr>
<td>7:30 AM – 8:30 AM</td>
<td>Breakfast (continental)</td>
</tr>
<tr>
<td>9:00 AM – 9:15 AM</td>
<td>WELCOME &amp; OVERVIEW • Jill Mesirov (ISC Board Member); and Mike Bevil and Joanne S. Luciano, Conference Co-chairs</td>
</tr>
<tr>
<td>9:15 AM – 10:00 AM</td>
<td>KEYNOTE 1 • SMAR Semantics for Clinical Healthcare Delivery • Dr. Isaac Kohane</td>
</tr>
<tr>
<td></td>
<td>Director, Children’s Hospital Informatics Program, Henderson Professor of Pediatrics and Health Sciences and Technology, Harvard Medical School (HMS), Co-Director, HMS Center for Biomedical Informatics, Director of the HMS Countway Library of Medicine, Cambridge, MA – USA</td>
</tr>
<tr>
<td>10:05 AM – 11:00 AM</td>
<td>LITERATURE</td>
</tr>
<tr>
<td></td>
<td>Executing Semantics Across Documents:</td>
</tr>
<tr>
<td></td>
<td>Bringing Science Into Context • Presenter: Anita de Waard</td>
</tr>
<tr>
<td></td>
<td>The Biospecimen Repository as Library:</td>
</tr>
<tr>
<td></td>
<td>How HeLa is like Moby Dick • Presenter: James McCusker</td>
</tr>
<tr>
<td>11:00 AM – 11:20 AM</td>
<td>Break</td>
</tr>
<tr>
<td>11:20 AM – 11:40 AM</td>
<td>TECH TALK 1 • Javascript – The Key to Successful SemWeb Deployments • Presenter: Jans Aasman, Franz Inc.</td>
</tr>
<tr>
<td>11:45 AM – 12:05 PM</td>
<td>TECH TALK 2 • Managing BigData® in Bioinformatics • Presenter: Bryan Thompson, SYSTAP LLC</td>
</tr>
<tr>
<td>12:05 PM – 1:15 PM</td>
<td>Lunch</td>
</tr>
<tr>
<td>1:15 PM – 2:00 PM</td>
<td>KEYNOTE 2 • E-Science Dictates E-Publication — Nanopublications as a Substrate for In-silico Knowledge Discovery • Dr. Barend Mons</td>
</tr>
<tr>
<td></td>
<td>Scientific Director, Netherlands Bioinformatics Center, Biosemantics Group Leader, Leiden University Medical Centre, Netherlands</td>
</tr>
<tr>
<td>2:05 PM – 3:30 PM</td>
<td>COMMUNITY/KNOWLEDGE MANAGEMENT</td>
</tr>
<tr>
<td></td>
<td>The VIVO Ontology: Enabling Networking of Scientists • Presenter: Ying Ding</td>
</tr>
<tr>
<td></td>
<td>Domain Knowledge and Provenance-Integrated Knowledge Organization System Represented with RDFS and SPARQL • Presenter: Young Soo Song</td>
</tr>
<tr>
<td></td>
<td>E-Diary Data Collection in Neurology and Psychiatry: Computational Achievements and Challenges • Presenter: Ron Calvanio</td>
</tr>
<tr>
<td>3:30 PM – 3:45 PM</td>
<td>Break</td>
</tr>
<tr>
<td>3:45 PM – 5:10 PM</td>
<td>DRUG DISCOVERY, LINKED DATA</td>
</tr>
<tr>
<td></td>
<td>Target Identification Using an Integrated Subset of the Yeast Interactome with Chemical Genomic Data in RDF • Presenter: Nadia Anwar</td>
</tr>
<tr>
<td></td>
<td>Chem2Bio2RDF: Linked Open Data for Drug Discovery • Presenter: Bin Chen</td>
</tr>
<tr>
<td></td>
<td>Using Linked Open Data to Inform the Drug Discovery Process • Presenter: James Snowden</td>
</tr>
<tr>
<td>5:15 PM – 6:00 PM</td>
<td>KEYNOTE 3 • Inside the Mind of Watson • Dr. Chris Welty</td>
</tr>
<tr>
<td></td>
<td>IBM Research Scientist, T.J. Watson Research Center New York, USA</td>
</tr>
<tr>
<td>6:00 PM – 6:05 PM</td>
<td>DAILY CLOSING REMARKS • Mike Bevil and Joanne S. Luciano, Conference Co-chairs</td>
</tr>
</tbody>
</table>
### Friday, February 24

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:30 AM – 10:00 AM</td>
<td>Registration</td>
</tr>
<tr>
<td>7:30 AM – 8:30 AM</td>
<td>Breakfast (continental)</td>
</tr>
<tr>
<td>8:30 AM – 8:45 AM</td>
<td>REVIEW PREVIOUS DAY • Mike Bevil and Joanne S. Luciano,</td>
</tr>
<tr>
<td></td>
<td>Conference Co-chairs</td>
</tr>
<tr>
<td>8:45 AM – 9:30 AM</td>
<td>KEYNOTE 4 • The Knowledge Reengineering Bottleneck • Dr. Rinke Hoekstra</td>
</tr>
<tr>
<td></td>
<td>Knowledge Representation &amp; Reasoning Group, VU University Amsterdam, Leibniz Center for Law, University of Amsterdam, Netherlands</td>
</tr>
<tr>
<td>9:30 AM – 10:25 AM</td>
<td>SEMANTICS IN RESEARCH</td>
</tr>
<tr>
<td></td>
<td>SDlink: An Integrated System for Linking Biological and Biomedical Semantic Data • Presenter: Alexandre Francisco</td>
</tr>
<tr>
<td></td>
<td>SPO: An Ontology for Describing Host-pathogen Interactions Inherent to Streptococcus Pneumoniae Infections • Presenter: Cátia Vaz</td>
</tr>
<tr>
<td>10:25 AM – 10:40 AM</td>
<td>Break</td>
</tr>
<tr>
<td>10:40 AM – 12:35 PM</td>
<td>PHARMACOVIGILANCE &amp; SAFETY</td>
</tr>
<tr>
<td></td>
<td>Dynamic Enhancement of Drug Product Labels Through Semantic Web Technologies • Presenter: Richard Boyce</td>
</tr>
<tr>
<td></td>
<td>Adverse Events Following Immunization: Standardization, Automatic Case Classification and Signal Detection • Presenter: Mélanie Courtot</td>
</tr>
<tr>
<td></td>
<td>Exploitation of Semantic Methods to Cluster Pharmacovigilance Terms • Presenter: Natalia Grabar</td>
</tr>
<tr>
<td></td>
<td>Annotation Analysis for Testing Drug Safety Signals • Presenter: Trish Whetzel</td>
</tr>
<tr>
<td>12:35 PM – 1:30 PM</td>
<td>Lunch</td>
</tr>
<tr>
<td>1:30 PM – 2:15 PM</td>
<td>KEYNOTE 5 • WolframAlpha and the Quest for Computational Knowledge • Dr. Stephen Wolfram</td>
</tr>
<tr>
<td></td>
<td>Founder &amp; CEO Wolfram Research, Wolfram Research, Champaign, IL – USA</td>
</tr>
<tr>
<td>2:20 PM – 3:15 PM</td>
<td>CLINICAL HEALTHCARE</td>
</tr>
<tr>
<td></td>
<td>Using Ontologies in the Age-Phenome Knowledge-base (APK) • Presenter: Eitan Rubin</td>
</tr>
<tr>
<td></td>
<td>Intelligent Surveillance of Health Care-associated Infections with SADI Semantic Web Services • Presenter: Christopher Baker</td>
</tr>
<tr>
<td>3:15 PM – 4:15 PM</td>
<td>PANEL DISCUSSION • Semantics for data: the role of the Semantic Web in orchestrating the new data driven world • Moderators: Mike Bevil and Joanne S. Luciano</td>
</tr>
<tr>
<td>4:15 PM – 4:30 PM</td>
<td>FUTURE ACTIONS • Mike Bevil and Joanne S. Luciano, Conference Co-chairs</td>
</tr>
<tr>
<td>4:30 PM</td>
<td>CONFERENCE ADJOURNS</td>
</tr>
</tbody>
</table>
KEYNOTE 1

Isaac (Zak) Kohane

SMArt Semantics for Clinical Healthcare Delivery

ABSTRACT: To Be Announced

BIOGRAPHY: Isaac (Zak) Kohane is the director of the Children’s Hospital Informatics Program and is the Henderson Professor of Pediatrics and Health Sciences and Technology at Harvard Medical School (HMS). He is also the co-Director of the HMS Center for Biomedical Informatics and Director of the HMS Countway Library of Medicine. Dr. Kohane leads multiple collaborations at Harvard Medical School and its hospital affiliates in the use of genomics and computer science to study diseases (particularly cancer and autism) through the perspective of biological development. He also has developed several computer systems to allow multiple hospital systems to be used as “living laboratories” to study the genetic basis of disease while preserving patient privacy. Among these, the i2b2 (Informatics for Integrating Biology and the Bedside) National Computing Center has been deployed at over 52 academic health centers internationally.

Dr. Kohane has published over 200 papers in the medical literature and authored a widely used book on Microarrays for an Integrative Genomics. He has been elected to multiple honor societies including the American Society for Clinical Investigation, the American College of Medical Informatics, and the Institute of Medicine. He leads a doctoral program in genomics and bioinformatics at the Division of Health Sciences and Technology at Harvard and MIT. He is also a practicing pediatric endocrinologist and father of three energetic children.
KEYNOTE 2

Barend Mons

E-Science Dictates E-Publication — Nanopublications as a Substrate for In-silico Knowledge Discovery

ABSTRACT: In a world of rapidly increasing complexity, interoperability of data is key and ‘data publishing’ will become the norm. New key technologies have altered our ability to generate massive data sets. And more data have been recognized as crucial data sources for research. Methods have been developed to mine massive data sets from different type of resources. These enormous strides in mining and measurement technologies have resulted in an exponentially growing flood of data: the ‘data explosion’ in the life sciences. The increasing complexity also asks for collaborative environments to master the data sets and turn them into insights. Life sciences projects and programmes are increasingly multidisciplinary and international. This brings many complexities in terms of collaboration and data, information and knowledge management far beyond ‘one’s own laboratory’. The transition to ‘e-science’ also dictates innovative ways to publish, share and cite valuable datasets. Barend will cover the recent development around nanopublication in the Life Sciences.

BIOGRAPHY: Barend Mons is a molecular biologist by training and received his PhD on genetic differentiation of malaria parasites from Leiden University. Subsequently he performed over a decade of research on malaria genetics and vaccine development in close collaboration with colleagues in developing countries. He served the research department of the European Commission in this field for 3 years as a seconded national expert and did gain further experience in science management at the Research council of The Netherlands (NWO). Barend was a co founder of three spin
off companies in biotechnological and semantic technologies.

In the year 200, Barend switched to the development of semantic technologies to manage big data and he founded the Biosemantics group. At present, Barend is Associate Professor in Bio-Semantics at the Department of Human Genetics at the Leiden University Medical Centre with an honorary appointment in the same discipline at the Department of Medical Informatics, Erasmus Medical Centre, University of Rotterdam, both in The Netherlands. He is also Scientific Director of the Netherlands Bioinformatics Center (NBIC), board member of the ELIXIR international board, an executive team member of the Open PHACTS IMI project and e-science integrator for the Life Sciences in the Netherlands e-Science Center.

His research is focused on nanopublications as a substrate for in silico knowledge discovery.
KEYNOTE 3

Chris Welty

Inside the Mind of Watson

ABSTRACT: Watson is a computer system capable of answering rich natural language questions and estimating its confidence in those answers at a level rivaling the best humans at the task. On Feb 14–16, in an historic event, Watson triumphed over the best Jeopardy! players of all time. In this talk I will discuss how Watson works at a high level with examples from the show.

BIOGRAPHY: Chris Welty is a Research Scientist at the IBM T.J. Watson Research Center in New York. Previously, he taught Computer Science at Vassar College, taught at and received his Ph.D. from Rensselaer Polytechnic Institute, and accumulated over 14 years of teaching experience before moving to industrial research. Chris’ principal area of research is Knowledge Representation, specifically ontologies and the semantic web, and he spends most of his time applying this technology to Natural Language Question Answering as a member of the DeepQA/Watson team and, in the past, Software Engineering. Dr. Welty was a co-chair of the W3C Rules Interchange Format Working Group (RIF), serves on the steering committee of the Formal Ontology in Information Systems Conferences, is past president of KR.ORG, on the editorial boards of AI Magazine, The Journal of Applied Ontology, and The Journal of Web Semantics, and was an editor in the W3C Web Ontology Working Group. While on sabbatical in 2000, he co-developed the OntoClean methodology with Nicola Guarino. Chris Welty’s work on ontologies and ontology methodology has appeared in CACM, and numerous other publications.
KEYNOTE 4

Rinke Hoekstra

The Knowledge Reengineering Bottleneck

ABSTRACT: Knowledge engineering upholds a long-standing tradition that emphasizes methodological issues associated with the acquisition and representation of knowledge in some (formal) language. This focus on methodology implies an ex ante approach: “think before you act”. The rapid increase of linked data poses new challenges for knowledge engineering, and the Semantic Web project as a whole. Although the dream of unhindered knowledge reuse is a technical reality, it has come at the cost of control. Semantic web content can no longer be assumed to have been produced in a controlled task-independent environment. When reused, Semantic Web content needs to be remoulded, refiltered and recurated for a new task. Traditional ex ante methodologies do not provide any guidelines for this ex post knowledge reengineering; forcing developers to resort to ad hoc measures and manual labour: the knowledge reengineering bottleneck.

BIOGRAPHY: Rinke Hoekstra is a researcher at the Knowledge Representation & Reasoning group of the VU University Amsterdam and at the Leibniz Center for Law of the University of Amsterdam, where he received his PhD on the topic of ontology representation and design patterns. Rinke’s main area of research is knowledge representation and engineering, specifically ontologies, the semantic web and linked data. He spends most of his time applying this technology to law and government information, and recently forayed into the domains of science and the humanities.

Rinke was a member of the OWL Working Group that developed OWL 2, and the eGov Interest Group of the W3C, and is editorial board member of the Semantic Web journal. He was program chair of OWLED 2009, and served on several program committees for conferences on AI and Law, and Semantic Web. Rinke is the main author of the LKIF Core ontology of basic legal concepts, and one of the initial developers of the MetaLex XML format for legal sources.
KEYNOTE 5

Stephen Wolfram

Wolfram|Alpha and the Quest for Computational Knowledge

ABSTRACT: To Be Announced

BIOGRAPHY: Stephen Wolfram is a distinguished scientist, inventor, author, and business leader. He is the creator of Mathematica, the author of A New Kind of Science, the creator of Wolfram|Alpha, and the founder and CEO of Wolfram Research.

Wolfram was educated at Eton, Oxford, and Caltech. He published his first scientific paper at the age of 15, and had received his PhD in theoretical physics from Caltech by the age of 20. In recognition of his early work in physics and computing, Wolfram became in 1981 the youngest recipient of a MacArthur Prize Fellowship.

In 1986 Wolfram founded the first research center and the first journal in the field, Complex Systems. Then, after a highly successful career in academia—first at Caltech, then at the Institute for Advanced Study in Princeton, and finally as Professor of Physics, Mathematics, and Computer Science at the University of Illinois—Wolfram launched Wolfram Research, Inc.

Wolfram has been president and CEO of Wolfram Research since its founding in 1987. In addition to his business leadership, Wolfram is deeply involved in the development of the company’s technology. He is personally responsible for overseeing all aspects of the functional design of the core Mathematica technology. He also directs the design and development of pioneering projects, such as Wolfram|Alpha, which was launched in 2009 as a long-term project to make as much of the world’s knowledge as possible computable, and accessible to everyone.

Wolfram has a lifelong commitment to research and education. In addition to providing software for a generation of scientists and students, Wolfram’s company maintains some of the web’s most visited sites for technical information. Wolfram is also increasingly active in defining new directions for education, especially in the science he has created.

http://www.stephenwolfram.com/about-sw/
Mashing HC and LS Data

Presented by

In part 1, you will be introduced to Semantic Web technologies, including RDF, RDFS, OWL, and SPARQL. You’ll also learn how to issue basic SPARQL queries from JavaScript.

In part 2, you will learn about Linked Data—Semantic Web data on the Web. This section will cover crawling and querying Linked Data, and then will take a look at health-related linked data such as DrugBank, DailyMed, SIDER, Diseasome, DBpedia, poss. LinkedCT, and PubMed. You’ll also look at mashing up linked data with personal health records and talk about some use cases and demo applications.

In part 3, you will build on what you’ve learned to accomplish these tasks:
• Convert a MAGE-TAB file from ArrayExpress to linked data and publish it using LODspeakr
• Add a summary graph to the experiment page to display the distribution of biospecimens by characteristic
• Load the data into GenePattern and extract differentially expressed genes
• Convert the gene list to RDF data cube and load into the triple store
• Create a heatmap from the gene list

Presenter:
Dominic DiFranzo, Rensselaer Polytechnic Institute
James McCusker, Rensselaer Polytechnic Institute
Joshua Shinavier, Rensselaer Polytechnic Institute

RPI Tutorial Coordinator: Joanne Luciano, Rensselaer Polytechnic Institute

CSHALS Tutorial Coordinator: Lee Feigenbaum, Cambridge Semantics
Executing Semantics Across Documents: Bringing Science Into Context

Anita de Waard, Elsevier Labs, Utrecht, Netherlands

ABSTRACT: In my presentation I will show how semantic technologies and Linked Data are forming the backbone of a new form of science publishing, where a paper is presented within three types of context. The first type of context is that of the research process. As we find better ways to integrate research data, executable components and workflow representations with the scientific narrative, we hope to add richness, depth and accountability to publications and improve the reader’s ability to evaluate and replicate the findings. The second type of context is that of the specific features of the object of study such as patient characteristics for clinical reports, species and subspecies for animal studies, or other experimental parameters such as instrumentation details. The third type of context we wish to enable the reader to have access to is the knowledge preceding and succeeding a given paper. By identifying the key claims the authors make and linking them to their supporting evidence both within and across papers, we hope to allow an infrastructure that will enable more straightforward ways of assessing trust and validity when assessing new information. I will demonstrate these principles with three use cases which we are working on together with our academic collaborations, pertaining to clinical guidelines, drug-drug interactions, and neuroscientific knowledge integration.

The Biospecimen Repository as Library: How HeLa is like Moby Dick

James McCusker, Rensselaer Polytechnic Institute, Troy, United States

ABSTRACT: Provenance-oriented data models are becoming critical for fostering interoperability among scientific workflow systems. Tools to manage laboratory systems and biorepositories record the actions of people and equipment in order to keep track of exactly what has happened to experimental artifacts. We explore the similarities between the library science standard Functional Requirements for Bibliographic Resources (FRBR) and
requirements for biospecimen management in research settings. Abstractive provenance, or the ability to describe entities and their history at multiple levels of abstraction, is used in FRBR to describe the relationship between a particular copy of a book and the concept of that book. A similar treatment can describe the relationship between a cell line, various physical colonies of cells from that cell line, and the originating organisms. We propose a similar standard, Functional Requirements for Biological Resources (FRBioR), to describe those requirements and an ontology that integrates with the W3C draft PROV provenance model and ontology.
Javascript — The Key to Successful SemWeb Deployments

Jans Aasman, Franz Inc. Oakland, CA, United States

Ideally, declarative query languages (ie SQL, SPARQL) would be so powerful that we would never need to perform any procedural server side programming in the database. However, the reality is that every Enterprise application backed by a relational database relies heavily on server side programming language link PL/SQL or a very vendor specific binding to Java/C++.

Currently, the W3C nor the Semantic Web community have a proposal for server side scripting languages or named services and from our view in the trenches we see that Javascript has all the right properties to be both this scripting language and the basis for named services.

During this presentation we will demonstrate our Javascript compiler and a Javascript library that can perform all the basic handling of RDF quads, indices, and databases. This server side Javascript can also include SPARQL and Prolog and has rich functionality for temporal and spatial functions and can be used to write graph based algorithms that will work at server speed.

MongoDB’s wide adoption has created a large demand for working with JSON objects. Our proposed library supports the MongoDB API to make working with JSON objects in combination with RDF nearly transparent. On top of all this, we can take any server defined program or function and make it instantly available as a REST call which instantly opens the Semantic Web to a host of programming talent.
Managing Bigdata® in Bioinformatics

Bryan Thompson, SYSTAP LLC. Greensboro, NC, United States

Bigdata® is a horizontally scaled open source semantic web database platform running on a single machine or a cluster. I will introduce the bigdata architecture, summarize some of its key differentiators and new features, including analytic query support, and show how different groups are using Bigdata® to tackle bioinformatics problems.

SYSTAP, LLC leads the development of the bigdata open source platform and offers consulting services related scalable information architectures and services and support for the bigdata platform. Bigdata® is available under both open source and OEM licenses.
The VIVO Ontology: Enabling Networking of Scientists

Ying Ding, Indiana University, Bloomington, United States

ABSTRACT: VIVO, funded by NIH, utilizes Semantic Web technologies to model scientists and provides federated search to enhance the discovery of researchers and collaborators across disciplines and organizations. VIVO ontology is designed with the focus on modeling scientists, publications, resources, grants, locations, and services. It incorporates classes from popular ontologies, such as BIBO, Dublin Core, Event, FOAF, geopolitical, and SKOS. VIVO data is annotated based on the VIVO ontology to semantically represent and integrate information about faculty research (i.e., educational background, publications, expertise, grants), teaching (i.e., courses, seminars, training), and service (i.e., organizing conferences, editorial boards, other community services). The VIVO ontology has been adopted nationally and internationally, and enables the national and international federated search for finding experts. VIVO is an open source Semantic Web application that, when populated with researcher interests, activities, and accomplishments, enables discovery of research and scholarship across disciplines and organizations. The VIVO core ontology models the academic community in order to provide an consistent and connected perspective on the research community to various shareholders, including students, administrative and service officials, prospective faculty, donors, funding agencies, and the public. The major impetus for NIH to fund the VIVO effort to “develop, enhance, or extend infrastructure for connecting people and resources to facilitate national discovery of individuals and of scientific resources by scientists and students to encourage interdisciplinary collaboration and scientific exchange” . The application is in use at the seven institutions of the NIH VIVO project and has been adopted or to be adopted by several other universities (e.g., Harvard University) and organizations in the USA (e.g., the United States Department of Agriculture), and several universities or institutions in Australia and China (e.g., Queensland University of Technology, Chinese National Academy of Sciences)
(Gewin, 2009). More specifically, VIVO can support discovering potential collaborators with complementary expertise or skills, suggesting appropriate courses, programs, and faculty members according to students’ interests, and facilitate research currency, maintenance and communication. For example, a Computer-Aided Drug Discovery (CADD) group may want to find and team up with a computer specialist and a group using in vivo experiments in drug discovery. If the VIVO core ontology is implemented in the hypothetical situation, the group leader can search across experts in computer science and molecular biology. In this paper, we present a relatively comprehensive discussion of the development of the VIVO core ontology, including the latest updates.

**Domain Knowledge and Provenance-Integrated Knowledge Organization System Represented with RDFS and SPARQL**

Young Soo Song, University of Alabama at Birmingham, United States

**ABSTRACT:** Objectives and Motivation: Although semantic web technologies are expanding as a framework to construct knowledge organization system (KOS), without controlling data flow based on rules and consensus, its adoption will be limited. We have previously addressed this problem by defining a Markov process for user operators associated with a KOS, S3DB. The idea was that by annotating existing assertions with the domain neutral S3DB tags, the user-operator states describing the provenance could then be tracked by a parallel algebraic process. That solution includes a mechanism for resolving both the merging and the migration of multiple, often conflicting, provenance. This mechanism is currently supported by a open source prototype (s3db.org) with a SPARQL endpoint and a query language, S3QL. Although the core concept of S3DB includes both domain knowledge and provenance model and its implementation is currently used in several institutions, they are loosely coupled because domain knowledge model was expressed as RDFS and provenance model as numeric computation. In this study we seek to bridge between the logic and
algebraic representations by describing user-operators as a RDFS model such that the integrated representation can be resolved by a SPARQL 1.1 engine.

**METHOD:** Semantics of S3DB provenance model were thoroughly analyzed and represented as a semantically equivalent SPARQL query. While S3DB domain knowledge model is a pure RDFS model, its provenance model is a numerical model, which receives arguments from asserted RDF triples and produces outputs as inferred triple through the successive procedures. In this model, asserted triples represent assigned user operator relationships between users and entities and new relationships between them are inferred through numerical procedures. Semantically reinforced SPARQL 1.1 can simulate these numerical procedures. In particular, propagation of user operators corresponded to SPARQL designed with property paths and merging to SPARQL designed with aggregation function. These phrases could be assembled into SPARQL having subquery as a part. Although the procedure of merging was performed during propagation of user operations in the original numerical model, our SPARQL model performed merging after propagation is completed, producing the same results while not being affected by the order of the procedures.

**RESULTS:** For the proof of concept, our integrated model was implemented with ARQ version 2.8.8., although any triple store or application supporting SPARQL 1.1 should deliver the same results. Provenance model was tested upon the cancer genome atlas (TCGA) data, containing microarray and sequencing data for over 500 cancer patients. Verification of the validity of our model needs only three steps, 1) installation of ARQ, or equivalent application supporting SPARQL 1.1, 2) importing of knowledge model and test TCGA data, and 3) executing of query representing our provenance model. Effective user operators inferred from the query could be stored in the other namespaces separate from the assigned user operators.

**CONCLUSIONS:** As a consequence, it is argued, complex provenance scenarios can be accommodated by data stores equipped with a
SPARQL endpoint. This result signifies that the proposed solution can be handled in a scalable and distributed manner by regular triple stores.

**E-Diary Data Collection in Neurology and Psychiatry: Computational Achievements and Challenges**

Ron Calvanio, Massachusetts General Hospital & Harvard Medical School, Cambridge, United States

Additional Authors: F. Buonanno, MD

Dr. Calvanio will present e-diary data recorded by patients undergoing outpatient treatment in a neurology clinic at the Massachusetts General Hospital. Patients had, or were suspected of having, a sudden onset disorder: a stroke, a traumatic brain injury, etc. Patient symptom complaints were: sensory or motor spells, headaches, emotional outbursts, fatigue, sleep disturbances, cognitive lapses, or odd behavior. Personalized e-diaries were designed to identify routine events that may have influenced symptom expression. Identification of symptom influences was then used to resolve diagnostic issues and to enhance treatment outcomes. Dr. Calvanio will show: 1) how e-diary data reveal symptom influence patterns, many of which patients are not aware; 2) how identification of these influences improves care; 3) what the computational challenges are in data coding, data analysis, and data pattern representation.
Target Identification Using an Integrated Subset of the Yeast Interactome with Chemical Genomic Data in RDF

Nadia Anwar, General Bioinformatics Reading, United Kingdom

ABSTRACT: Semantic web technologies provide a well-established, efficient and cost-effective data integration strategy. We demonstrate the advantages these technologies offer in target discovery through combining experimental evidence from several target discovery technologies using the linked data paradigm. Drug discovery methods such as drug induced HaploInsufficiency profiling (HIP) are here combined with other chemical genomic data and genetic interaction networks to improve the sensitivity and specificity of target identification. We demonstrate the value of this integration using the yeast interactome with complementary experimental evidence within a fungal target discovery pipeline for triaging of hit lists, target identification and target deconvolution. Genetic interaction profiles generated from yeast deletion strains, using methods such as SGA, describe the relationships between genes. These profiles integrated into a network of genetic interactions are used to uncover and predict the functions of uncharacterized genes. Chemical Genomic data describes the influence of small molecules on biological systems and are used to characterize the effect of compounds at the cellular level. Chemical genomic methods including HaploInsufficiency Profiling (HIP), Homozygous profiling (HOP) and Multi-copy Suppression Profiling (MSP) are commonly used together to overcome the limitations in individual technologies. For example, HIP is used to identify small molecule (drug) targets, however, HIP is limited to molecules that inhibit cell growth and will fail to identify targets with functional paralogs. While HIP identifies direct targets, HOP is especially useful for providing insight into potential drug interactions. A combination of these approaches provides a more complete view, specifically identifying both on-target effects and off-target effects. Since HIP, HOP and MSP are based on the same principles, a combined approach, although more time consuming and expensive, delivers more comprehensive data. An alternative
combined approach is based on the idea of re-using established genomic scale data. Costanzo et al. re-use their clustered genetic interactions (GI) by correlating these with chemical-genetic interactions (CGI) from HOP. They successfully re-use their GI and CGI data and demonstrate that their combined profiles are complementary to HIP. Following Costanzo et al, we used semantic technologies to integrate the genetic interaction profiles described in their paper, with a test set of compounds assayed using HIP. Specifically, we have followed the linked data approach using clustered networks of GI and CGI in yeast, layered over chemical genomic experiments. We show that this method is not only a cost-effective integration strategy for these data but it also simplifies the discovery of the target as well as relevant interactions. Creating a foundational resource of these data in this fashion allows new experimental results and clusters to be layered into the network efficiently, moving from boutique, case by case integration, to a scalable and robust integration resource. We demonstrate how this integrated data set can be used to identify profiles for compounds of interest, its target, and aid the visualization of the targets network proximal to the compound’s immediate target. Finally, we demonstrate how providing such a comprehensive view of view of the data eases the investigation of the compound’s mechanism of action.

Chem2Bio2RDF: Linked Open Data for Drug Discovery

Bin Chen, Indiana University Bloomington, United States

ABSTRACT: A critical barrier in current drug discovery is the inability to utilize public datasets in an integrated fashion to fully understand the actions of drugs and chemical compounds on biological systems. There is a need for not only a resource to intelligently integrate the heterogeneous datasets pertaining to compounds, drugs, targets, genes, diseases, and drug side effects now available, but also robust, effective network data mining algorithms that can be applied to such integrative data sets to extract important biological relationships. In this talk, we discuss (i) the creation of
the Chem2Bio2RDF for drug discovery data, integrating chemical compounds, protein targets, genes, metabolic pathways, diseases and side-effects using Semantic Web technologies, and (ii) the development of innovative data mining algorithms to facilitate drug discovery. Chem2Bio2RDF incorporates 25 public datasets related to systems chemical biology, grouped into 6 domains: chemical (PubChem Compound, ChEBI, PDB Ligand); chemogenomics (KEGG Ligand, CTD Chemical, BindingDB, MATADOR, PubChem BioAssay, QSAR, TTD, DrugBank, ChEMBL, Binding MOAD, PDSP, PharmGKB); biological (UNIPROT, HGNC, PDB, GI); systems (KEGG Pathway, Reactome, PPI, DIP); phenotypes (OMIM, Diseasome, SIDER, CTD diseases); and literature (MEDLINE/PubMed). The number of RDF triples is approximately 110 million. We developed the domain ontology (called Chem2Bio2OWL) to better integrate these 25 datasets. The primary classes of this ontology are: SmallMolecule, MacroMolecule, Disease, SideEffect, Pathway, BioAssay, Literature and Interaction based partially on the BioPAX classes. The primary classes were further refined in accordance with current instance data structure. We proposed and tested several graph mining and machine learning algorithms (e.g., Bio-LDA, path finding, subgraph mining and diversity ranking) on the generated Chem2Bio2RDF linked open dataset to facilitate drug discovery. We found that our Bio-LDA model used the bio-terms, journal information and word information to characterize the topic providing a better representation of topics than the simple LDA model, which only can provide the word representation. Rosiglitazone is one of several thiazolinediones on the market for diabetes. Our path finding algorithm presents the set of most informative and diverse associations between the drug and the potential side effects, which shows different causes of the hepatitis side effect. Our constraint-based subgraph and diversity ranking algorithm can detect the inhibition of Catechol O-methyltransferase (COMT) in Parkinson’s disease. By combining information from Drugbank, Pubchem and Uniprot, we can find information regarding the gene that Tolcapone and Entacapone targets, its name, the protein it encodes,
Pubmed articles related to their interaction with COMT, and the structure of the protein it targets. In this talk, we demonstrated the potentials of data mining and graph mining algorithms to identify hidden associations that could provide valuable directions for further exploration at the experimental level. In the future, we will focus on using the identified associations and paths existing between various bio terms to predict the potential connection of other unknown bio terms.

**Using Linked Open Data to Inform the Drug Discovery Process**

James Snowden, UCB Celltech Slough, United Kingdom

**ABSTRACT:** The treatment of disease and identification of new targets via which the symptoms / causes of disease can be treated is one of the cornerstones of drug discovery research in the pharmaceutical sector. Whilst much of the information for these areas is available, it is distributed in many systems both internally and externally. Therefore, the main issue with gathering the required information is actually one of time / resource. In response to this at UCB, we have developed the Target Information Platform (TIP) and Disease Information Platform (DIP) systems to collate key information relating to targets and disease respectively and make this available in a single portal for easy access by our scientists. This approach is underpinned by the capabilities provided by semantic technology and in particular Linked Open Data (LOD), which allows complete querying of available data sources in a quick and automated manner. The public LOD system which is comprised of SPARQL endpoints over key biological data sources is queried using SPARQLMOTION scripts through the TopBraid composer system. This takes in a single data item (Uniprot ID for target, disease name for disease) which is used to pull data an initial endpoint (UniProt / Diseaseome). The results from this are parsed and where relevant, additional calls are made out to endpoints for other data sources. The end result of this is that a RDF data package is generated which collates together relevant information
from multiple sources in the public domain. Additionally for DIP, literature, patent and omics data is queried and stored in a Triplestore. Web pages are generated from this information which are provided to the scientists. The key benefits that have been derived through this approach so far have been speed, completeness of data searching and increasing the availability of target / disease information. A target search that may have taken 2-3 days work for 1-2 scientists can now be done in 5-10 minutes. This frees up scientist time, provides target information faster and allows many more targets to be queried. The data searching is done in a standardised manner with the aspect of human error removed and also more consistency in terms of data returned for targets and disease. Finally, providing the information returned in a central portal means that scientists always know where to go to access the information. All of these benefits are in some way related to the semantic / LOD approach used. Disadvantages of this approach are mostly related to technical issues of endpoint uptime / availability and also updating of information within the endpoints. This work has demonstrated that it is possible to utilise the public LOD framework in an automated manner that exemplifies that linked data principle by starting from a single point of information to gather detailed data. It has returned information relating to key concepts of vital importance to drug discovery that have helped to optimise this process at UCB and has demonstrated practical utility for semantics and LOD.
SDlink: An Integrated System for Linking Biological and Biomedical Semantic Data

Alexandre Francisco, Technical University of Lisbon, Portugal

ABSTRACT: Nowadays, with the decreasing cost and increasing availability of high-throughput technologies, an enormous amount of biological and biomedical data is becoming available. Such data is usually represented and stored in different formats and platforms, most of the times off line and not standardized. The automatic integration of data from different databases suffers from several caveats, the most notable being the lack of interfaces for automatic querying and running integration and analysis tools. In order to solve some of these issues, semantic technologies have been proposed and used with great success. In these work we propose an integrated environment for querying, retrieving and analyzing linked data, suitable for users unfamiliar with such technologies, solving an issue that has been deterring a more generalized adoption of semantic methodologies in biology and biomedicine.

METHOD: The new sdlink system (http://kdbio.inesc-id.pt/sdlink) assumes that data is annotated following a given ontology and provides data views, including graphical representations, and a friendly querying interface. The querying interface was developed to be used by semantic technologies unfamiliar users, where one can for instance define a query by means of a point and click simple interface, which is then translated to SPARQL. The sdlink system uses Virtuoso OSE as the underlying triplestore. To address user concerns with respect to security and privacy, the system supports user/project control access, based on OpenID for authentication and FOAF+WAC for authorization. The system is being used by two FP7 European projects, with good results both in what concerns scalability and usability by non-expert users. We made also available a public project for evaluation purposes (http://kdbio.inesc-id.pt/sdlink/lubm/).

RESULTS: Our results were twofold. First, through the development and deployment of sdlink, we were able to use semantic technologies and linked data on two large projects were most...
people were unaware of these technologies or of any reason to use them. The main contribution was an interface that simultaneously allows users to retrieve and query linked data, and does not lose expressiveness, efficiency or scalability. In particular, the system is self-adaptable to ontology changes and data transformations, depending only on the update of underlying ontologies. The projects where the system was tested are dealing with heterogeneous data, including sequence data and experiment results, resulting from several teams and work packages, that in the end became integrated, browsable and queryable. The data stored comprises about one million triples, which can be queried in less than one second for most usual queries.

**CONCLUSIONS:** The development of sdlink, and its deployment in a real scenario, allowed us to concluded about the importance and usefulness of semantic technologies, namely for domain representation and data integration. More importantly, it was possible to show that, by developing suitable interfaces, any user can benefit from such technologies. Currently, the unfriendliness of most semantic technologies, in particular in the fields of biology and biomedicine, have struggle the adoption of these technologies. The sdlink system is proposed to overcome this problem and to bring semantic technologies and linked data to a broader audience.

**SPO: An Ontology for Describing Host-pathogen Interactions Inherent to Streptococcus Pneumoniae Infections**

Cátia Vaz, Poly Institute of Lisbon, Portugal

**ABSTRACT:** Over the past twenty years, the study of infection has tended to consider individual virulence factors or host factors. The Pneumopath project (www.pneumopath.org/), a FP7 European research project, has the objective of studying the host-pathogen interactions during infection of Streptococcus pneumoniae and finding new targets for diagnosis and treatment. This research purports to identify the most important and consistently involved host and pneumococcal factors, in contrast to previous approaches,
where factors were studied in isolation. The transmission of Streptococcus pneumoniae to a new host can result in asymptomatic colonization or progress to invasive disease. The infection can be determined by multiple attributes of both host and pathogen, being important to take into account the epidemiological and genomic characterization of pneumococcal strains, the results from experiments that evaluate host or pneumococcal responses to infection or different environmental challenges, and also the results from experiments that identify host genetic susceptibility factors. In this work we propose Spo (http://kdbio.inesc-id.pt/~cvaz/pneumopath/), an ontology developed in the context of the Pneumopath project, which provides terms and semantic constructs for annotating all aspects of host-pneumococcal interactions.

**METHOD:** The data considered includes the characterization of pneumococcal strains, typing information, as well as data of in vitro and in vivo experiments with animals and cell models, relevant for identifying new targets to combat pneumococcal diseases. Some of these data are scattered across numerous information systems and repositories, each with its own terminologies, identifier schemes, and data formats. The need to share such data brings challenges for both data management and annotation, such as, the need to have a common understanding of the concepts that describes host-pneumococcal interactions. Thus, semantic annotation and interoperability become an absolute necessity for the integration of such diverse biomolecular data. Moreover, given the heterogeneous environment inherent to the project, the ontology construction took into consideration contributions from all partners, leading to a well-grounded set of concepts and annotations.

**RESULTS:** Spo provides a framework to represent mentioned host-pneumococcal interactions, being flexible enough to accommodate the rapid changes and advancement of research and achieve data interoperability and interchange. This has been only possible because of semantic Web recommended practices for clearly specifying names for things and relationships, expressing data using standardized and well-specified knowledge representation
languages. The ontology described in OWL Lite v1.0 includes 36 classes, 24 object properties and 43 data properties.

**CONCLUSION:** The main contribution of this work was not only Spo, but all the approach and methodology for its construction in the context of a large research project, where many people were not aware of semantic technologies. The proposed ontology does not only describe knowledge in this field, but also allows for validating and aggregating existing knowledge, which is essential for data integration. Furthermore, the ability to accurately describe the host-pneumococcal interactions through the use of Spo has facilitated the implementation of information systems capable of coping with the heterogeneous types of data and, by using well known semantic technologies, it allowed users to query data and discover new knowledge.
Dynamic Enhancement of Drug Product Labels Through Semantic Web Technologies

Richard Boyce, University of Pittsburgh, United States

ABSTRACT: FDA-approved drug product labeling (packages insert or PI) is a major source of information intended to help clinicians prescribe drugs in a safe and effective manner. Unfortunately, drug PIs have been identified as often lagging behind the drug knowledge expressed in the scientific literature, especially when it has been several years since a drug has been released to the market. Out-of-date or incomplete PI information can increase the risk of otherwise preventable adverse drug events. This can occur directly if the PI fails to provide information that is needed for safe dosing or to properly manage drugs known to interact. Clinicians might also be indirectly affected if they depend on third party drug information sources, and these sources fail to add information that is available in the scientific literature but not present in the PI. We are creating a Linked Data store that will enable the drug PI to be expanding as new information becomes available in the scientific literature. The goal of the Linked Data store will be to provide clinicians, patients, and the maintainers of drug information resources with the most complete and up-to-date information on particular claims made within a PI. We are focusing on 25 currently-marketed psychotropic medications (nine antipsychotics, twelve antidepressants, and four sedative hypnotics). To construct this Linked Data repository, we aim to use Natural Language Processing (NLP) technologies identify core claims in the scientific literature and various web-based data sources that pertain to pharmacokinetic drug-drug interactions, age-related changes in clearance, metabolic clearance pathways, and genetic polymorphisms that can affect metabolism. This work aligns with the CSHAL themes “Linked Data”, “Text Analysis, NLP; Question Answering”, “Data Modeling: Ontologies, Taxonomies”, and “Clinical Applications”.

METHOD: We will identify the core rhetorical components of the content sources using a basic Scientific Discourse ontology constructed (and compatible with) biomedical discourse ontologies
(i.e., SWAN, OAC and AO) and discourse annotation metadata (specifically CoreSC). The ensuing discourse annotations will distinguish between facts, hypotheses, and evidence statements, and will be automatically recognised in text following an information extraction approach similar to conceptualisation zoning. The expected result is a Linked Open Data Node, a Triple store and a SPARQL endpoint available for use by different patient, clinician, and pharmacoepidemiology-centered data sources. Human readable summaries will also be generated to expand on existing PI information.

**RESULTS:** While we are in the early planning phases of the project, we have built a prototype system that demonstrates the concept by identifying how claims on metabolic clearance and drug-drug interactions could be updated in two drug PIs with evidence from the scientific literature.

**CONCLUSIONS:** We envision using the resulting Linked Data store as the back end for a system that provides pharmacokinetic information on age-related clearance changes, metabolic clearance pathways, pharmacokinetic drug-drug interactions, and genetic polymorphisms. After developing a demonstrator for the 25 psychotropics, we anticipate that it will be feasible to subsequently deploy our system for any given drug.

**Adverse Events Following Immunization: Standardization, Automatic Case Classification and Signal Detection**

Mélanie Courtot, British Columbia Cancer Research Centre, Vancouver, Canada

**Additional Authors:** Ryan R. Brinkman BC Cancer Agency & University of British Columbia, Vancouver, Canada; Alan Ruttenberg University at Buffalo, NY, United States

**ABSTRACT:** Analysis of spontaneous reports of Adverse Events Following Immunization (AEFIs) is an important way to identify potential problems in vaccine safety and efficacy and summarize experience for dissemination to health care authorities. However, current reporting methods are not sufficiently controlled. While there is general adoption of Medical Dictionary of Regulatory
Activities (MedDRA) in the reporting systems we consider, definitions are not provided for MedDRA terms, reports are not annotated in a consistent manner, differing in experience of annotator, and annotation is done either at entry time, or post-hoc. Sometimes, only the final adverse event code is saved, discarding evidence supporting the diagnosis. Because of these practices, interpretation of such spontaneous reports is tedious, costly and time consuming. The Adverse Event Reporting Ontology (AERO) we are building plays a role in increasing accuracy and quality of reporting, ultimately enhancing response time to adverse event signals.

**METHODS:** In order to address these deficiencies, we work with the Brighton Collaboration who has done extensive work towards standardization of case definitions and diagnostic criteria for vaccine adverse events. Based on our initial results with AERO, a working group has been established within the Brighton network, including representation from the Public Health Agency of Canada (PHAC) and the US Food and Drug Administration (FDA), to incorporate logical representations of Brighton case definitions into AERO, with the aim of increasing quality and accuracy of AEFI reporting. As an example, only 9% of the Vaccine Adverse Event Reporting System (VAERS) anaphylaxis reports post-H1N1 vaccination early 2010 were correctly annotated with the MedDRA anaphylaxis term. Working within the framework being established by the Open Biological and Biomedical Ontologies (OBO) Foundry, the Adverse Events Reporting Ontology (AERO) first documents assessments of relevant signs and symptoms textually. These elements of AEFI reports are then logically defined by being positioned into a hierarchy and related to each other in a way that supports computing an overall diagnosis. Our system allows automatic inference of a diagnosis according to the Brighton criteria based on the evidence encoded in the MedDRA annotations. As an additional test of our approach we will also attempt to parse the textual section of VAERS reports and annotate them with AERO terms with the aim of using the logic encoded in AERO to determine diagnoses as defined in the Brighton Guidelines.
Our approach allows us to unambiguously refer to a specific set of carefully defined signs and symptoms at the time of data entry, as well as an overall diagnosis that remains linked to its associated signs and symptoms. The adverse event diagnosis is formally expressed, making it amenable to further querying for example for statistical analysis (“what percentage of patients presented with motor manifestations?”) and at different levels of granularity. Finally, by enabling automatic processing of adverse events reports, we will decrease time and money needed for their evaluation. This may allow earlier detection of adverse events signal in the datasets, and trigger a warning for experts to further investigate.

Exploitation of Semantic Methods to Cluster Pharmacovigilance Terms

Natalia Grabar, Universite Lille Villeneuve d’Ascq, France

ABSTRACT: Pharmacovigilance activity is related to the collection, analysis and prevention of adverse drug reactions (ADRs) likely to be caused by drugs. This activity is achieved thanks to the case reporting to the pharmacovigilance authorities and pharmaceutical industries. Before their inclusion in pharmacovigilance databases, the ADRs of these case reports are coded with terms from dedicated terminologies, such as MedDRA. The analysis of the collected ADRs is related to the safety surveillance within these databases. It relies on the identification of relations between a drug and an ADR. It has been observed that some couples {drug, ADR} are not activated, when they should be. The main cause then is that MedDRA is a fine-grained terminology and that the encoding of the adverse reactions with MedDRA may have an impact on the signal dissolution: similar and close ADRs may be encoded with different terms and during the analysis of the databases they will remain isolated and the safety risk detection may be underestimated.

METHODS: We propose to exploit semantic resources and methods provided by Natural Language Processing and by Computer
Sciences for automatic generation of clusters of the MedDRA terms, which have close semantic and clinical meaning. We exploit the ontological resource ontoEIM and MedDRA terms. The SMQs are exploited as the gold standard. Among the methods, we use semantic distance approaches, lexically-based methods for detection of hierarchical and synonymy relations between terms, as well as several clustering methods. The obtained clusters of terms are compared with the existing SMQs, both hierarchical and non-hierarchical. The results are evaluated with three metrics: precision, recall and f-measure. Results are evaluated quantitatively (against the gold standard) and qualitatively (by medical and pharmacovigilance experts).

RESULTS: Various factors have been tested: exploitation of formal definitions, several semantic distance approaches, weighting of the semantic axes within formal definitions, clustering methods, comparison and combination of semantic distance and lexical methods. We obtain results which indicate that the generated clusters can assist the creation of new SMQs or the hierarchical structuring of terms within SMQs. Depending on the SMQs, we obtain interesting results with the semantic distance approach (precision between 36% and 87%, recall between 15% and 77%) and for the lexical approach (precision between 10% and 92%, recall between 3% and 33%). Moreover, these two methods provide complementary results. Indeed, safety topics are better modeled with one or another of the methods. For instance, the generation of the Agranulocytosis cluster has poor results with semantic distance approach: the relevant terms are spread within the ontoEIM resource. Although this grouping shows high performances with the lexical method: the relevant terms have semantic similarities which can be detected at the lexical level.

CONCLUSION: The performed experiences indicate that it is possible to generate meaningful clusters of terms on new safety topics in order to assist the creation of new SMQs. The exploited methods can also be exploited for the refinement of the hierarchical structure of the existing SMQs.
Annotation Analysis for Testing Drug Safety Signals

Trish Whetzel, Stanford University, United States

ABSTRACT: Introduction R is used versus using coded data alone. Changes in biomedical science, public policy, and electronic health record (EHR) adoption have converged recently to enable a transformation in health care. While analyzing structured EHRs have proven useful in different contexts, the true richness and complexity of health records—roughly 80 percent—lies within the free-text clinical notes and it is crucial to develop methods to test for drug safety signals throughout the EHR. Using ontology-based approaches, we computed the risk of having a Myocardial infarction (MI) on taking Vioxx for Rheumatoid arthritis (RA) using the annotations created on the textual notes for over 1 million patients in the Stanford Clinical Data Warehouse (STRIDE).

METHODS: Based on the NCBO Annotator Web service, we created a standalone NCBO Annotator Workflow that is highly optimized for both time and space. The workflow was extended to incorporate negation detection, the concept recognizer Unitex, and uses ontologies from BioPortal. To reproduce the risk of MI following Vioxx treatment, we identified patients in STRIDE with a pattern of RA, who are taking Vioxx, and then suffer MI. To identify patients with RA and MI, we scanned through structured data of 25 million coded ICD9 diagnoses for codes beginning with the ICD9 codes for RA and MI. We also scanned through the normalized annotations of the unstructured data, to look for non-negated mentions of MI and RA. We denote the first occurrence or mention of the condition as t0(RA) and t0(MI). We did not have access to the structured medication data; therefore, we relied upon annotations derived from the textual notes to identify patients taking Vioxx. We scanned through the normalized annotations of the unstructured data to look for non-negated mentions of Vioxx or rofecoxib. We denote the first occurrence or mention of the drug as t0(Vioxx).

RESULTS: The Annotator Workflow was enhanced in both time and space and processed 9.5 million patient notes in 7 hours using 4.5 GB of disk space. From the observed patient counts, we
constructed a contingency table and obtained a reporting odds ratio (ROR) of 2.058 with a confidence interval (CI) of [1.804, 2.349] and proportional reporting ratio (PRR) of 1.828 with CI of [1.645, 2.032]. The uncorrected \( \chi^2 \) statistic was significant with a p-value < 10\(^{-7} \). In comparison, without using the unstructured data and only using the ICD9 coded data, the results were more ambiguous. The corresponding risks for the results without the unstructured data were: ROR=1.524 with CI=[0.872, 2.666] confidence interval; and PRR=1.508 with CI=[0.8768, 2.594]; and \( \chi^2=0.06816 \).

**CONCLUSIONS:** We have significantly scaled the NCBO Annotator Workflow to computationally annotate the free-text narrative of over 9.5 million reports from STRIDE. Our results demonstrate that unstructured data in the EHR provide a viable source for testing drug safety signals using annotations created from the textual notes. Our analysis recapitulated the latent Vioxx risk signal and found that the risk is far more perceptible when ontology-based analysis methods of unstructured data in the EHR is used versus using coded data alone.
Using Ontologies in the Age-Phenome Knowledge-base (APK)

Eitan Rubin, Ben Gurion University Beer Sheva, Israel
Additional Author: Nophar Geifman, Ben Gurion University, Beer sheva, Israel

ABSTRACT: The importance of age in biomedical research and clinical care has resulted in an abundance of publications linking age and phenotypes. However, these data are organized such that searching for age-phenotype relationships is prohibitively difficult. Recently, we described the Age-Phenome Knowledge-base (APK), a computational platform for storage and retrieval of information concerning age-related phenotypic patterns. Here we present and discuss the incorporation and use of ontologies and standardized vocabularies in the APK.

METHODS AND RESULTS: The Age-Phenome Knowledge-base contains evidence, such as scientific publications and clinical data analysis, connecting specific ages or age groups and phenotypes such as diseases. It makes extensive use of ontologies and fixed vocabularies in order to describe ages, diseases and other forms of phenotypes. Ages and age groups are described using the Age Ontology, a simple ontology developed for this purpose and based on the description of age-ranges in the Medical Subject Headings (MeSH). The Disease Ontology (DO) is used in APK to represent diseases while other forms of phenotypes are described by a subset of the Unified Medical Language System (UMLS) Metathesaurus. Complex searches are made possible by abstracting over the Age Ontology and the Disease Ontology’s hierarchical structures.

CONCLUSIONS: APK provides an example of how ontologies can be used in rapid development of new knowledge models. It makes integral use of ontologies and vocabularies to represent diseases and age groups in a standard, unambiguous way. Furthermore, the use of ontologies allows abstraction, which in turn makes it easy to develop/conduct complex queries.
Intelligent Surveillance of Health Care-associated Infections with SADI Semantic Web Services

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

ABSTRACT: Objectives and Motivation: Clinical Intelligence (CI) tools support data analysis for the purposes of clinical research, surveillance and rational health care management. Ad-hoc querying of clinical data is one desirable type of functionality. Since most of the data is currently stored in relational form, ad-hoc querying is problematic as it requires specialized technical skills and the knowledge of particular data schemas. A possible solution is semantic querying where the user formulates queries in terms of domain ontologies that are much easier to navigate and comprehend than data schemas. Existing approaches to semantic querying of relational data, based on declarative semantic mappings from data schemas to ontologies cannot cope with situations when some computation is required in query time. We are reporting preliminary progress on a project dedicated to the use of SADI Semantic Web services for semantic querying of clinical data for the surveillance of hospital-acquired infections (HAI).

METHOD: We implement semantic access to a Relational DB by using an ontology for HAI and modeling the RDB in it. The modeling is implemented by SADI Semantic Web services that can be automatically discovered and invoked based on the needs of a particular query. The main services draw data from the DB, but services bringing data from external resources are also used. Users formulate SPARQL queries using primitives from the ontology and execute them via a SADI query engine. The querying can be both ad-hoc and self-service because the users need not know RDB programming.

RESULTS: To test our approach in a CI scenario dedicated to the surveillance for HAI, we are prototyping a SADI-based infrastructure for semantic querying of The Ottawa Hospital datawarehouse (see, e. g.,). Our infrastructure includes an ontology defining concepts suitable for reasoning about Hospital-Acquired Infections and a number of SADI services on the datawarehouse.
To test the infrastructure, we write SPARQL queries representing questions a HAI surveillance professional would like to ask, such as “Which patients were diagnosed with SSI while they were taking corticosteroids?” or “How many diabetic patients were diagnosed with SSI?”. To facilitate temporal comparisons required by many competency questions, we created a time ontology and wrote a set of SADI services implementing temporal reasoning.

**CONCLUSIONS:** The main conclusion from our work on semantic querying so far is that the use of SADI services via a SPARQL interface is a viable general direction. Our approach will add to the pool of existing practical methods for semantic querying of RDB, at least in CI.

**PANEL DISCUSSION**

**FRIDAY, FEBRUARY 24**  3:15 p.m. – 4:15 p.m.

**Semantics for data: the role of the Semantic Web in orchestrating the new data driven world**

*Water water everywhere and not a drop to drink,* so goes the “rime of the ancient mariner.” As ever cheaper and widely available biomolecular screening methodologies flood a myriad of method specific data centers, filling that glass with task specific context is getting both harder and more rewarding. The role of the semantic web of linked data and web services to orchestrate cloud based resources, from drug development and biomedical research to personalized health care, appears poised to take center stage.
**Poster 1 • SDlink: An Integrated System for Linking Biological and Biomedical Semantic Data**

**Presenter:** Alexandre Francisco, INESC-ID / IST, Technical University of Lisbon, Portugal  
**Additional Authors:** Pedro Reis, INESC-ID / IST, Technical University of Lisbon, Portugal; Dário Abdulrehman, INESC-ID / IST, Technical University of Lisbon, Portugal; Cátia Vaz, INESC-ID / ISEL, Poly Inst of Lisbon, Portugal; Mauro Santos, INESC-ID / IST, Technical University of Lisbon, Portugal; Ana Freitas, INESC-ID / IST, Technical University of Lisbon, Portugal  
See page 28 for abstract from oral presentation.

**Poster 2 • SPO: An Ontology for Describing Host-pathogen Interactions Inherent to Streptococcus Pneumoniae Infections**

**Presenter:** Cátia Vaz, INESC-ID / ISEL, Poly Institute of Lisbon, Portugal  
**Additional Authors:** Pedro Reis, INESC-ID / IST, Technical University of Lisbon, Portugal; Alexandre Francisco, INESC-ID / IST, Technical University of Lisbon, Portugal; Susana Vinga, INESC-ID, Lisbon, Portugal; Ana Freitas, INESC-ID / IST, Technical University of Lisbon, Portugal  
See page 29 for abstract from oral presentation.

**Poster 3 • Chem2Bio2RDF: Linked Open Data for Drug Discovery**

**Presenter:** Bin Chen, Indiana University, Bloomington, United States  
**Additional Authors:** Ying Ding, Indiana University, United States; Philip Yu, University of Illinois, Chicago, United States; Eric Gifford Pfizer, United States; David Wild, Indiana University, United States  
See page 24 for abstract from oral presentation.

**Poster 4 • The VIVO Ontology: Enabling Networking of Scientists**

**Presenter:** Ying Ding, Indiana University, Bloomington, United States  
**Additional Authors:** Stella Mitchell, Cornell University, United States; Jon Corson-rikert, Cornell University, United States; Brian Lowe, Cornell University, United States; Bing He, John Hopkins University, United States  
See page 19 for abstract from oral presentation.
BioPAX, Biological Pathway Exchange, is an OWL ontology modelling biological pathway data. Biological pathways are constructs that biologists use to represent relationships between and within chains of cellular events. For example, metabolic pathways typically represent flow of chemical reactions, while signal transduction pathways represents the chain of interactions that transmit external signals received by a cell to deliver some response within the cell. These data are as heterogeneous as the numerous data sources (pathguide.org) that supply the data. Exchange, integration and annotation of these data is a considerable challenge. BioPAX was developed to ease the access, use, exchange and aggregation of pathway data. This poster will highlight the recent community developments. The current specification, BioPAX Level 3 and its supporting API, PaxTools, were released by the community in 2009. Since this release, the BioPAX community has focused on supporting developers with the transition from L2 to L3, community organisation, interoperability with other standards and future directions from user feedback. In 2010, BioPAX joined forces with SBML, SBGN and other standards, to form the ‘C0mputational Modeling in B1ology’ NEtwork (COMBINE). This initiative aims to co-ordinate the development of the various community standards and formats. Through learning from the experiences of community organization in other successful standards, the BioPAX community have re-orgnaised themselves. In place now is an invited Scientific Advisory Board and an elected Editorial Committee who are now co-ordinating governance and proposal development with the COMBINE netowrk. The annual BioPAX meetings are also now co-ordinated with the COMBINE network, providing economy of scale in development of standards through a shared forum to share experiences, and enabling the standardization efforts to work co-operativley. These meetings are
organized into two events a Hackathon (Harmony May 21–25, 2012 in Masstricht) and the COMBINE forum (Combine August 15-19 2012, Toronto). The BioPAX community is also responding to feedback from a survey undertaken in 2011. Community members, consumers and data providers gave valuable information on how they use BioPAX, the difficulties they faced and how they want to see the specification progress in the future. This feedback will be used by the new governance teams to help establish the specification in the areas it is currently used, to help extend the community beyond current usage and also to determine future directions for the community. To get involved or find out more about BioPAX see www.biopax.org join the mailing list biopax-discuss@googlegroups.com or attend a meeting in 2012.

Poster 6 • Using ontologies in the Age-Phenome Knowledge-base

Presenter: Eitan Rubin, Ben Gurion University, Beer sheva, Israel
Additional Authors: Nophar Geifman, Ben Gurion University, Beer sheva, Israel
See page 39 for abstract from oral presentation.

Poster 7 • Dynamic Enhancement of Drug Product Labels through Semantic Web Technologies

Presenter: Richard Boyce, University of Pittsburgh, PA, United States
Additional Authors: Jodi Schneider, Digital Enterprise Research Institute, Ireland; Michael Taylor, Microsoft, United States; Maria Liakata, EBI, United Kingdom; Anita De Waard, Elsevier, United States
See page 32 for abstract from oral presentation.

Poster 8 • A Case Study in Using Literature to Find Predicate Relationships and Indirect Associations

Presenter: James Dixon, Linguamatics Ltd., Newton, MA, United States
Additional Authors: David Milward, Linguamatics, United Kingdom
OBJECTIVES AND MOTIVATION: Gene expression has been the focus of much research, especially for treatment of carcinomas. Recently attention has turned to smaller RNA molecules that are involved in post-transcriptional regulation, microRNAs. MicroRNAs (miRNAs) are known to bind to complementary sequences on
target messenger RNA transcripts (mRNAs). MiRNA-expression profiling of different neoplasms has identified signatures associated with diagnosis, staging, progression, prognosis and response to treatment. In addition, profiling has been exploited to identify miRNA genes that might be involved in cancer or oncogenic pathways. To obtain a better insight into the connection between miRNAs and diseases requires understanding of the relationships between miRNAs and genes, and the relationship between the relevant genes and diseases. This paper compares and links together data from different sources: algorithmic predictions, experimental evidence and text mined literature.

**METHOD:** There are a number of publically available databases that have the miRNA to gene mapping, usually via statistical calculations. However, few have established the mechanism of action. Each mechanism is different and may matter to a researcher. Using the Linguamatics I2E text mining platform we were able to mine research literature (Medline abstracts) using natural language processing (NLP) to add relational information to the miRNA-gene combinations. A particular challenge was the nomenclature for miRNAs, which may include prefixes and suffixes. For example, they may be prefixed to distinguish species, such as hsa-miR-19a for human and mmu-miR-19a for mouse. Our approach treated them as a single family since in general, their function is very similar. This also allowed us to extract literature results where the species is not identified and only the family name is used. Since our interests lie in connecting miRNA to carcinomas, we also used the same literature source, Medline, to extract gene to carcinoma relationships, to allow linking between the miRNA and the diseases via the genes they affect.

**RESULTS:** Using I2E, we found over 6000 miRNA to gene relationships from Medline abstracts. These relationships overlapped to some extent with commonly used databases in the genomic field, for instance TargetScan(1004), TarBase(135) and miRecords(316). The overlap of all three databases to each other was similar to what was found with the I2E results. Focusing on
a single carcinoma, non-small cell lung cancer (NSCLC), as an example, we were able to extract over 400 indirect relationships between miRNA and NSCLC, where other public databases had less than 50 miRNA to NSCLC associations.

**Conclusions:** Since all of the public database information used had modest overlap with the results from the literature, we are confident that we added not only relational information to the miRNA-gene interactions, but also added novel relationships to the miRNA-disease connections. In addition, we have extended our network from miRNA to gene and gene to disease to a more interesting relationship of miRNA to disease via their indirect links. Creating these associations will provide researchers new avenues to explore, lead to new target identification, and hopefully, new disease treatments.

**Poster 9 • Image Retrieval in Controlled English**

**Presenter:** Tobias Kuhn, Yale University School of Medicine, New Haven, CT, United States  
**Additional Authors:** Michael Krauthammer, Yale University, United States

The Yale Image Finder (YIF) project aims at improving biomedical image and document retrieval by developing advanced image parsing and indexing strategies. To this end, we have deployed a YIF search engine, which allows for keyword searches against indexed Pubmed Central open access images. Authors often follow well-accepted layouts when depicting experiment results as gels, graphs or plots, and use image text in an equally structured fashion for labeling different image elements. Image text placement often conveys higher-level semantics, such as the names of proteins being studied under different experimental conditions. We are currently exploring innovative ways for allowing YIF users to access such structured image text content. Here, we propose the use of a controlled language interface that guides users in composing natural language queries (“Find an image where X is measured under the condition Y”) that are be subsequently mapped to indexed image text content. Our approach is based on controlled natural language, i.e. a restricted subset of English with a precise and unambiguous mapping to logic. We present a prototype called Rice (Retrieving
Images through Controlled English) that is based on an interface we developed for a different domain (annotated text corpora) and adopted for image mining. Users can write seemingly natural queries like “Find an image that is a Western blot and where ‘p38’ is compared to ‘MKK3’” which is subsequently translated into a logical representation like “western-blot & compared(p38,MKK3)”. Such logical representations can then be matched with the formal model that we extract from images found in biomedical papers. One serious problem with controlled natural language is that it is very easy to read and understand but hard to write. Our prototype solves this problem by providing a predictive editor, with which users construct syntactically correct sentences in an iterative and guided way. For any partial sentence, the predictive editor of Rice shows the possible continuations in the form of different menu boxes. In this way, users do not need to know about the restrictions of our language beforehand. Previous evaluation has shown that this editor is very easy to use after very little or no training. Typical users of search engines are not familiar with logic notations and rarely have the time to learn one. Existing query interfaces are either very simple (i.e. keyword-based) or too complex to be usable without training. With Rice, complex queries can be written in a natural and intuitive way. The interface should be immediately accessible to researchers interested in the results represented in images of the biomedical literature. Rice supports queries with directed relationships “… where A is measured under the condition B”, resulting in the retrieval of highly specific image sets. In contrast, keyword searches cannot build such refined query representations, and cannot easily tell apart a related query “… where B is measured under condition A”. Our prototype is still incomplete, but we believe that it nicely demonstrates the potential of our approach, and the positive results of previous work make us confident of its practicality.
The cell cycle is an essential, highly conserved, complex process. Understanding the cell cycle is important in understanding development, aging, and the progression of many diseases including cancer. Mouse Genome Informatics (MGI) is the international database resource for the laboratory mouse, providing genetic, genomic, and biological data to facilitate the study of the mouse as a model for human health and disease. We have recently developed a mouse cell cycle ontology as a novel approach to data integration for the diverse data on the laboratory mouse that is available at MGI. Currently at MGI, 1070 mouse genes are functionally associated with the cell cycle and have been annotated to the Gene Ontology (GO) term ‘cell cycle’ and its descendants. This mouse cell cycle gene set also has a large body of additional experimental annotation: 8126 experimental GO annotations in addition to ‘cell cycle’; 581 genes have phenotypic alleles with 31,134 phenotype annotations describing 10,129 affected anatomical systems; 512 genes have curated OMIM (Online Medelian Inheritance in Man) associations to mouse models; 58 genes have pathway (MouseCyc) annotations; and 1055 genes have human orthologs. Many of these data are described by different ontologies from the Open Biomedical Ontologies (OBO): gene product function data is annotated using the Gene Ontology; mouse phenotype data using the Mammalian Phenotype Ontology; expression data using the Adult Mouse Anatomical Dictionary and the Edinburgh Mouse Atlas for embryonic stages. Our mouse cell cycle ontology provides a view across these distinct ontologies providing a richer description of the data. The analysis of data related to cell cycle processes requires an integrated view that pulls together as much data as possible. Our approach adapts and extends a method that has been used by other groups to develop cell cycle ontologies.
for other organisms, including human, yeast, and Arabidopsis. In this work, we describe the structure and content of our mouse cell cycle ontology, Mouse_CCO, as an ‘application’ ontology built on experimental evidence-based annotations for the specific purpose (application) of studying the cell cycle. The structure of Mouse_CCO provides the generic template for the ontology, which is then populated using 1070 mouse cell cycle genes along with all their annotations from MGI and several additional data resources. The data drives the structure and allows a user to ‘discover’ connections. As an experimental evidence-based ontology, it is particularly important to keep the ontology up to date. The two newly developed tools also described in this work simplify maintenance of the ontology: the first allows a user to download mouse genes and selected annotations in OBO format that is then used by the second tool, Oort (OBO Ontology Release Tool), to perform MIREOT-like procedures to create a merged ontology bringing in subsets of external ontologies. The final product is Mouse_CCO in both OBO and OWL formats that can be queried and explored using a variety of free, publicly available tools. Our hope is that this resource will facilitate hypothesis generation based on the cell cycle as a biological system.

**Poster 11 • OpenBEL, the BEL Framework, and the BEL Portal**

**Presenter:** Julian Ray, Selventa, Cambridge, MA, United States  
**Additional Authors:** Ted Slater, Selventa, United States; Natalie Catlett, Selventa, United States; David De graaf, Selventa, United States

The Biological Expression Language (BEL) and supporting technology platform, the BEL Framework, will be released by Selventa, in conjunction with Pfizer, to the life sciences community in Q1 of 2012. BEL and the BEL Framework are designed to promote the collection, sharing, and interchange of structured knowledge within and among organizations. The BEL Portal, at http://belframework.org/, is the online community home for BEL and the BEL Framework. The Biological Expression Language (BEL) is a language for representing scientific findings in the
life sciences in a computable form. BEL is designed to represent scientific findings by capturing causal and correlative relationships in context, where context can include information about the biological and experimental system in which the relationships were observed, the supporting publications cited, and the process of curation. BEL is intended as a knowledge capture and interchange medium, supporting the operation of systems that integrate knowledge derived from independent efforts. The BEL Language has been designed and used by our scientists and our customers for almost a decade. The language has been specifically designed to help scientists record life science facts in a way that is intuitive, easy to learn, concise, and appealing. A good language should help the user articulate an idea in a manner that is unambiguous, terse, and conveys the facts and associated contexts without loss or ambiguity. BEL is designed to do just this for life science applications. The current version of the language is small, which makes it easy to learn. BEL supports both causal and correlative relationships as well as negative relationships, which makes it suitable for recording a variety of experimental and clinical findings, and it can be used with almost any set of vocabularies and ontologies, which makes it highly adaptable and easy to adopt. BEL can be easily extended to annotate findings with use-specific contexts such as experimental and clinical parameters. The BEL Framework is an emerging open-platform technology specifically designed to overcome many of the challenges associated with capturing, integrating, and storing knowledge within an organization, and sharing the knowledge across the organization and between business partners. The BEL Framework provides mechanisms for knowledge capture and management; integration of knowledge from multiple, disparate knowledge streams; knowledge representation and standardization in an open, use-neutral format; creating customizable, computable biological networks from captured knowledge; and quickly enabling knowledge-aware applications using standardized application programming interfaces (APIs) across all major development platforms. Registering on the BEL Portal gives you access to more detailed documentation about BEL and the BEL Framework,
and also allows you to participate in our community section and offer your views, opinions, and suggestions on the language and framework as well as keeping you informed on the progress of the official launch. Once you register you will have access to example documents, best practices, technical specifications, configuration guides, code examples, a wiki, and discussion groups.

**Poster 12 • Semantic Integration to Characterize Microbial Pathogens: Multi-resource Enrichment of Experimental Proteomic and Genomic Datasets**

**Presenter:** Erich Gombocz, IO Informatics, Inc., Berkeley, CA, United States  
**Additional Authors:** James Candlin, Sage-n Research, United States

Bacterial and viral-caused infectious diseases account for major health threats globally, yet the characterization, identification and understanding of them has been scientifically challenging. This is mainly due to the fact that while there is a wealth of information (and even complete genomes) available, its integrated utilization in context of the biological system to better understand causes and similarities in infectious diseases is still in its infancy. This poster tries to address some of the many obstacles involved in this endeavor as it attempts to identify peptides from different microorganism with common mechanism of actions causing disease, and to use them as biomarkers to detect pathogenic microbial threats prior to onset of disease symptoms to help in outbreak prevention. The presented workflow to accomplish this goal consist of 5 steps. The first step is a thorough peptide analysis of microorganism via mass spectrometry and their identification by sequence scoring (Sorcerer, indexed SEQUEST search, BioWorks). The second step is the annotation of peptides with genes and genomic sequences relevant to protein expression to qualify the accuracy of the identification. Step 3 involves the use of public domain microbial databases (PATRIC, ICTV, VIDA, Viral ORFeome, miRBase) to semantically integrate the experiments with organism taxon-specific functional genomic and pathway information relevant to diseases caused by the pathogens. Based on sequence similarity, sequences are clustered into homologous
protein families (HPFs), and those protein families are enriched with annotations including functional classification, related protein structures, taxonomy, protein length, boundaries of conserved regions and bacterial or virus-specific genes. Further enrichment is achieved through addition of disease-related pathways (BioCyc, KEGG). The resulting knowledgebase provides a network with functional annotations to peptides and their relationships to diseases (Sentient Knowledge Explorer). In Step 4, those peptides in the network are identified which have similar disease-causing functions and appear in several pathogens. Interrogating the network via semantic queries (SPARQL) results in discovery of key pathway intersections commonly involved in the disease. The last step is the creation of molecular marker signatures (SPARQL, Applied Semantic Knowledgebases - ASK) and test their validity as decision support in multiplexed assays. Future applications will apply this technology for rapid detection of biological threats, to characterize origin and type of disease outbreaks and to develop preventive measures (such as broadly applicable drugs or vaccines) effective for entire classes of pathogenic organism.

**Poster 13 • The Quad Economy of a Semantic Web Ontology Repository**

**Presenter: Trish Whetzel, National Center for Biomedical Ontology, Stanford, CA, United States**

**Additional Authors:** Paul R. Alexander, Stanford University, United States; Mark A. Musen, Stanford University, United States; Natalya F. Noy, Stanford University, United States

BioPortal is an open library of biomedical ontologies that can be accessed using a Web-based user interface or RESTful Web services. The Web-based user interface allows users to browse, search, and visualize ontologies and facilitates community participation in the ontology lifecycle, including reviews of ontologies, mappings between terms, comments and new term proposals. A suite of Web services, including Web services that expose information about terms in ontologies, mappings, notes, and metadata about the ontologies themselves, drives the Web-based interface. The NCBO Web services provide a common XML output of ontology content regardless of the ontology...
representation format, however there is no single uniform storage for the ontologies and their metadata. As the amount of information in BioPortal and number of hits to the NCBO Web services increase, a more scalable solution is needed. To address these issues, we analyzed the use of a quad store since quad stores easily scale to millions of triples and provides SPARQL query access to the ontologies. Currently each ontology in BioPortal includes the materialization of all owl:imports. Thus, if a small ontology imports a large ontology then the former becomes a large ontology. Taking into account that BioPortal stores multiple versions of an ontology, the problem is reproduced for every version. Our hypothesis was that we could optimize the number of quads in the system using a more granular model where owl:imports are not materialized and every ontology graph contains its own RDF triples without the triples from the owl:imports ontologies. One of the questions to be answered is the optimization ratio—in number of triples—when using an ontology-per-graph model versus a closure-materialized model. Of the 149 OWL ontologies reviewed, there are 299 ontologies in the import closure (i.e., if we follow all the owl:imports links from the 149 ontologies, we will create a set of 299 ontologies). These 299 OWL ontologies contain 303 owl:imports, the materialized import closure is a set of 495 owl:imports. We also reviewed the number of re-used triples. Ontologies with no imports gather 5.4M triples in the system; ontologies with one import 1.7M; ontologies with 2-9 imports reach 0.5M triples; and more than 10 imports 2.1M. To conclude, our analysis shows that while ontology reuse is still far from being the norm, effective reuse is a goal worth pursuing and the level of reuse can have significant implications for the scalability of ontology storage systems.
Posters

Poster 14 • A PubMed Search Engine for Rat Genome Curation at RGD

Presenter: Weisong Liu, The Medical College of Wisconsin, Milwaukee, WI, United States
Additional Authors: Mary Shimoyama, The Medical College of Wisconsin, United States; Melinda Dwinell, The Medical College of Wisconsin, United States; Howard Jacob, The Medical College of Wisconsin, United States

The Rat Genome Database (RGD) is a collaborative effort between leading research institutions involved in rat genetic and genomic research. One of the main tasks of RGD is to curate rat gene related literatures and enter the information into our database. In this work, we built a PubMed search engine to help our curators locate paper-of-interest more efficiently. Using NCBI’s Entrez Utilities for Java, we have created a pipeline to weekly download PubMed data in XML format. We parse the XML files using a parser generated from NCBI’s efetch_pubmed.xsd file to extract information such as PMID, title, abstract, publication date and authors. The parsed information is stored in a MySQL database. This makes it easier for us to further utilize the information. By making use of the GATE and the UIMA frameworks, we built another pipeline to extract ontology (gene ontology, rat strain ontology, disease ontology, sequence ontology and organism ontology) terms and synonyms, and gene names/symbols from the PubMed titles and abstracts stored in the database. Some third-party plugins, such as Abner, OrganismTagger and MetaMap, were also used in this pipeline. The output of this pipeline includes ontology IDs, term positions within an abstract, and matching types. In order to make our framework scalable, we set up a small Hadoop cluster. The XML files are compressed and stored on Hadoop HDFS. Using the MapReduce framework, we can run our XML parsing and information extraction pipeline in many parallel threads. This dramatically reduced the total processing time comparing to a single-threaded program. The pipeline can also run on Amazon Web Services’ Elastic MapReduce. Along with the stored PubMed information, the pipeline output is fed into a Solr server. All information is indexed by Apache Lucene. With a web based user-interface, a user can search for PubMed abstracts by entering PMIDs, authors,
Posters

publication dates, terms, ontologies, ontology IDs, gene names and(or) gene symbols. The search results are ranked by relevance. Matched terms are sorted by frequency of appearance in an abstract.

**Poster 15 • The Biospecimen Repository as Library: How HeLa is like Moby Dick**

** Presenter: James McCusker, Rensselaer Polytechnic Institute, Troy, United States**

See page 15 for abstract from oral presentation.

**Poster 16 • Turning Biological Knowledge into Mathematical Models, Automated**

**Presenter: Oliver Ruebenacker, University of Connecticut Health Center, United States**

**Additional Authors:** Michael Blinov, University of Connecticut Health Center, United States; Ion Moraru, University of Connecticut Health Center, United States; James Schaff, University of Connecticut Health Center, United States

Living organisms are so enormously complex that we need computer simulations to understand the consequences of their vast biochemical reaction networks. As we uncover an increasing part of these networks, our established knowledge is increasingly stored in free web databases and available for query and download in machine-readable formats, especially in the RDF/OWL-based community standard Biological Pathways Exchange (BioPAX). The available data is massive and growing, e.g. Pathway Commons stores 1,700 pathways, 414 organisms, 440,000 interactions and 86,000 substances. This data is fully linked with open controlled terminologies such as gene ontology (e.g. anatomical features) and other free online databases such as ChEBI (chemicals), KEGG (genes a.o.), UniProt (proteins) and PubMed (publications). Automatic use of this knowledge for computer simulations of biological organisms has been an ongoing challenge. Now, Systems Biology Pathway Exchange (SBPAX), a BioPAX extension, allows the inclusion of quantitative data and systems biology terms, especially the Systems Biology Ontology (SBO). SBPAX support has been implemented by the Virtual Cell, Signaling Gateway Molecule Pages and System for the Analysis of Biochemical Pathways - Reaction Kinetics (SABIO-RK). For the first time, a mathematical model can be automatically built and fully annotated from a pathway of interest.
Storing biomedical data in various structured forms, like biomedical ontologies, and at different locations have brought about many challenges for answering queries about the knowledge represented in these ontologies. One of the challenges is to represent a complex query in a natural language and get its answers in an understandable form. Another challenge is to answer complex queries that require appropriate integration of relevant knowledge stored in different places and in various forms, and/or that require auxiliary definitions, such as, chains of drug-drug interactions, cliques of genes based on gene-gene relations, similarity/diversity of genes/drugs. Furthermore, once an answer is found for a complex query, the experts may need further explanations about the answer. We have built a software system, called BIOQUERY-ASP, that handles all these challenges. **METHOD:** We have addressed the three challenges described above using the high-level knowledge representation formalism and efficient automated reasoners of Answer Set Programming (ASP) - a declarative programming paradigm that supports various semantic Web technologies. To address the first challenge, we have developed a controlled natural language for biomedical queries about drug discovery; this language is called BIOQUERY-CNL. Then we have built an intelligent user interface that allows users to enter biomedical queries in BIOQUERY-CNL and that presents the answers (possibly with explanations or related links, if requested) in BIOQUERY-CNL. To address the second challenge, we have developed a rule layer over biomedical ontologies and databases that not only integrates the concepts in these knowledge resources but also provides definitions of auxiliary concepts. We have introduced an algorithm to identify the relevant parts of the rule layer and the knowledge resources with respect to
the given query, and used automated reasoners of ASP to answer queries considering these relevant parts. To address the third challenge, we have developed an intelligent algorithm to generate an explanation for a given answer, with respect to the query and the relevant parts of the rule layer and the knowledge resources. The overall system architecture for BIOQUERY-ASP is presented in the figure included in the supporting document.

**RESULTS:** We have shown the applicability of BIOQUERY-ASP to answer complex queries (specified by experts) over large biomedical knowledge resources.

**Poster 18 • Proposed Ontology for Seizure and Epilepsy**

**Presenter:** Robert Yao, Arizona State University, United States  
**Additional Authors:** Graciela Gonzalez, Arizona State University, United States; Jeffrey Buchhalter, Phoenix Children’s Hospital, United States

The understanding and classification of seizures and epilepsy syndromes have constantly changed with the advent of new knowledge from new technologies. Ontologies provide a structured knowledge framework that could aid in more precisely defining and standardizing terminologies and diagnoses. This in turn could enhance the abilities of researchers and clinicians to pinpoint the causes of a disorder, discover new treatment measures, and improve patient outcomes.

**HYPOTHESIS:** We hypothesize that a more refined ontology for seizures and epilepsy syndromes that adequately reflects the latest measurements, observations and medical findings can be used to assist empirical diagnosis of epilepsy and to potentially differentiate new syndromes in a logical and standardized format.

**METHODS:** A review of previously proposed Seizure and Epilepsy classifications is being done to determine the most general way to classify each seizure, syndrome, and epilepsy. By analyzing and defining the building blocks of Epilepsy, an Epilepsy Ontology is iteratively formalized using Protege. Each seizure and syndrome will be instantiated to the ontology to determine if it provides a reasoning framework on epilepsy knowledge.
RESULTS: A poly-axial ontology is being defined to encode the conceptual building blocks of seizures and Epilepsy. The ontology will be open for both qualitative and quantitative evaluation when the data/evidence is available in preference over consensus expert opinion.

CONCLUSIONS: The aim of this ongoing work is to help clinicians better understand the etiology of seizures and definitions of and relationships between seizures and epilepsy syndromes, and to provide a more helpful path towards research, diagnosis, and treatment of the disorder. Eventually, this ontology could be expanded for use with other diseases, providing more structured definitions. Such a standard framework could also help pinpoint knowledge deficits which in turn should drive laboratory and clinical experiments to discover missing knowledge.

Poster 19 • Exploiting Ontology Information for Extracting Keyphrases from Biomedical Articles

Presenter: Kyu-Baek, Hwang Soongsil University, Seoul, Korea
Additional Authors: Sun Gon Kim, University of Seoul, Korea; Eunok Paek, University of Seoul, Korea

Keyphrases (or keywords) of a document serve a role of compactly representing its content. They can be used for indexing or summarization purposes. Our method for keyphrase extraction is based on supervised machine learning combined with ontology information. It consists of two stages: (1) keyphrase candidate generation and (2) keyphrase selection. In the first stage, keyphrase candidates are generated by extracting every unigram, bigram, and trigram of the words in the title and abstract of each article. Also, a set of ontology terms are assigned to each article. For this, any automated methods for ontology term assignment, e.g., vector space models, can be adopted. Ontology terms are used for expanding the set of keyphrase candidates. In specific, keyphrases, frequently co-occurred with the ontology terms assigned to a document, are added to its candidate set. In the second stage, keyphrases are selected from the expanded candidate set by supervised machine learning. Features for supervised learning include term and inverse document frequencies, length, first/last occurrence positions, and relationships with ontology terms. The confidence and
lift of an association rule between keyphrases and ontology terms are used for representing their relationships. Because multiple ontology terms are usually assigned to a document, ontology-related feature values are averaged across all of them.

RESULTS: The proposed method was applied to a dataset consisting of 1,799 articles from three journals in the biomedical literature, i.e., IEEE/ACM Transactions on Computational Biology and Bioinformatics, Journal of Computational Biology, and Journal of Proteome Research. The MeSH (Medical Subject Heading) descriptors, which constitute a biomedical ontology, are manually assigned to the articles published in these journals for PubMed indexing. These MeSH descriptors represent the subject content. In our experiments, MeSH descriptors were automatically assigned to each document of our dataset by a vector space model-based method. In addition, each article of these journals is annotated with about four to six author-provided keyphrases. These author keyphrases were used as a gold standard for keyphrase extraction evaluation. We conducted a 10-fold cross validation experiment using several supervised machine learning methods including naïve Bayes classifiers and Bayesian networks. The experimental results showed that the inclusion of ontology information improved the keyphrase extraction performance about 100% in terms of the F1-measure. When the number of extracted keyphrases was set to five, our method achieved an F1-measure of about 0.185 and the performance increase was 129%. We also compared our method with KEA, a method for keyphrase extraction using syntactic features (which is accessible at www.nzdl.org/Kea/index.html). Our method was always better than KEA regardless of the number of extracted keyphrases (the performance increase was from 2 to 98%). These results confirm the fact that semantic information about document topics plays a central role in keyphrase extraction.

CONCLUSIONS: We proposed a method for keyphrase extraction from documents using ontology information. Through a set of experiments, we showed that the inclusion of ontology information about document topics could greatly improve the performance in keyphrase extraction.
**Celebrating Twenty Years**

**ISM B 2012**

**Long Beach CA USA**

**SIGS & TUTORIALS JULY 13–14**

**CONFERENCE JULY 15–17**

---

**Plan to Attend!**


An Official Conference of the International Society for Computational Biology

---

**Register Early and Win!**

Register by May 11, 2012 and enter to win 4 nights accommodation at the Hyatt Regency Long Beach hotel for your conference stay.

---

**Keynote Speakers**

- **Andréj Sali**
  Univ. of California
  San Francisco, USA

- **Barbara Wold**
  California Institute of Technology
  Pasadena, USA

- **ISM B 20th**

- **Anniversary Keynote**
  **Richard H. Lathrop**
  Univ. of California
  Irvine, USA

- **ISM B 20th Anniversary Keynote**
  **Lawrence Hunter**
  Univ. of Colorado
  School of Medicine
  Denver, USA

- **ISCB Fellow Keynote**
  **Richard Durbin**
  Wellcome Trust Sanger Institute, Cambridge, UK

- **2012 ISCB Accomplishment By A Senior Scientist Award**
  **Gunnar von Heijne**
  Stockholm Univ., Sweden

- **2012 ISCB Overton Prize**
  **Ziv Bar-Joseph**
  Carnegie Mellon Univ., Pittsburgh, USA

---

**Conference Chairs**

- **Sydney Brenner**, Honorary Conference Chair,
  Howard Hughes Medical Institute, Ashburn, USA; Salk Institute, La Jolla, USA

- **Terry Gaasterland**, Conference Co-chair
  University of California San Diego, USA

- **Richard H. Lathrop**, Conference Co-chair
  University of California Irvine, USA

- **Burkhard Rost**, Conference Co-chair
  Technical University Munich, Germany