Christopher J.O. Baker  Greg Butler
Igor Jurisica (Eds.)

Data Integration in the Life Sciences

9th International Conference, DILS 2013
Montreal, QC, Canada, July 11-12, 2013
Proceedings

Springer
Preface

DILS was established in 2004 as a new bioinformatics workshop focusing on topics related to data management and integration. Now in its ninth year the conference continues to attract researchers from across a range of disciplines all of whom recognize the challenges faced by life scientists in managing and reusing data. Stakeholders involved in digital ecosystems and data ownership are able to generate large volumes of high-quality data and want to publish it to the widest possible audience for prospecting by scientists. And yet, data are not knowledge. The real value is the translation of the data into actionable knowledge. The methodologies and frameworks we depend on to facilitate this translation are still evolving, and the challenges in data management and reuse have grown rather than diminished over the last decade and are common across many disciplines.

Life science remains one of the leading domains and continues to create massive amounts of diverse data needing validation, curation, and annotation with meaningful descriptions and formatting according to open standards to ensure it is sharable between interoperable distributed systems and accessible by end users. Practitioners are, however, continually experimenting and the forum for discussing which methodologies have succeeded, which new technologies are now being adopted, for which particular tasks, and how they are used to integrate data for subsequent bioinformatic analysis is DILS.

This year, DILS received 23 papers to the main research track (both long and short papers). Four papers were accepted unconditionally. A further six were accepted with the provision that authors revised their papers in accordance with reviewers’ comments and provided detailed and itemized responses. All papers were subsequently verified by the Program Committee (PC) Chair and General Chairs.

Accepted papers cover a range of important topics including: algorithms for ontology matching, interoperable frameworks for text mining using Semantic Web services, pipelines for genome-wide functional annotation, automation of pipelines providing data discovery and access to distributed resources, knowledge-driven querying-answer systems, prizms, nanopublications, electronic health records and linked data. This year we opted to also offer an Early Career and Systems Track at the DILS workshop. At the time of writing, papers submitted to each track were still under review. These papers are not published in the research track proceedings.

DILS 2013 featured two keynote speakers. Firstly, Dr. Erich Gombocz, co-founder and CSO of IO Informatics a decade ago, is a veteran in applying systems biology approaches to pharmaceutical and clinical decision making based on semantic data integration and knowledge management technologies. Dr. Gombocz
presented the rationale for rethinking old problems, retooling with new methodologies and revisiting the process models that underpin our existing knowledge discovery in pharma and clinical practice in healthcare. Specifically, he advocates new patient-centric, precision-medicine healthcare models to change how drugs are developed, how trials are performed, and how patients are treated. His manuscript is included in the proceedings. Our second keynote speaker was Dr. Paolo Ciccarese, Assistant in Neuroscience at Massachusetts General Hospital and Instructor in Neurology at Harvard Medical School. He is known for his pioneering work on the Annotation Ontology, an RDF model for exchanging annotation. In his talk, Dr. Ciccarese introduced annotation as a form of “micro-integration,” in which typed, versioned, and provenance links are assigned between text and schema, text and data, or data and data. He showed how the Open Annotation standard facilitates both short- and longer-term data integration efforts, transforming content into smart and connected data.

DILS 2013 was held at Concordia University in Montreal, Canada, and was organized as part of a series of three co-located events known as the Semantic Trilogy. The two co-located events were the 4th International Conference on Biomedical Ontology and the 4th Canadian Semantic Web Symposium.

As the event co-chairs and editors of this volume, we would like to thank all authors who submitted papers, as well as the PC members and additional referees for their excellent work in evaluating the submissions. Special thanks go to Concordia University for providing us with the facilities to run the event, and the Semantic Trilogy organization team. Finally, we would like to thank Alfred Hofmann and his team at Springer for their cooperation and help in putting this volume together.

May 2013

Christopher J.O. Baker
Greg Butler
Igor Jurisica
Organization

General Chairs
Christopher J.O. Baker University of New Brunswick, Saint John, Canada
Greg Butler Concordia University, Montreal, Canada

Program Committee Chair
Igor Jurisica University of Toronto, Canada

Program Committee
Adam Lee University of Maryland, USA
Adrien Coulet Loria - INRIA Nancy Grand Est, France
Amar K. Das Stanford University, USA
Artjom Klein University of New Brunswick, Canada
Asif M. Khan National University of Singapore, Singapore
Bastien Rance NIH, Bethesda, USA
Brad Malin Vanderbilt University, USA
Christian Schönbach Kyushu Institute of Technology, Japan
David De Roure Oxford e-Research Center, UK
Dietrich Rebholz-Schuhmann EBI, UK
Erhard Rahm University of Leipzig, Germany
Fatima Al-Shahrour Broad Institute of MIT and Harvard, USA
Fleur Mougin University of Bordeaux Segalen, France
Guo-Qiang Zhang Case Western Reserve University, USA
Hasan Jamil University of Idaho, USA
James Cimino NIH/CC/OD, USA
Jörg Hakenberg Arizona State University, USA
Jerven Bolleman Swiss Institute of Bioinformatics, Switzerland
Jong Park Korea Advanced Institute of Science and Technology, Korea
Karen Eilbeck University of Utah, USA
Karin Verspoor National ICT, Australia, NISTA
Lawrence Hunter University of Colorado, USA
Marco Masseroli Politecnico di Milano, Italy
Marco Roos LUMC, University of Amsterdam, The Netherlands
Maria Esther Vidal Universidad Simón Bolívar, Venezuela
Matthew Hindle Synthetic and System Biology, Edinburgh, UK
VIII Organization

Michael Krauthammer  
Mong Li Lee  
Neil Sarkar  
Nigam Shah  
Paolo Missier  
Paolo Romano  
Peter Mork  
Radhakrishnan Nagarajan  
Rainer Winnenburg  
Satya Sahoo  
Tammy Cheng  
Vasant Honavar  

Yale University, USA  
National University of Singapore, Singapore  
University of Vermont, USA  
Stanford Center for Biomedical Informatics Research, USA  
Newcastle University, UK  
Institute of Genoa, Italy  
Noblis, USA  
University of Kentucky, USA  
NIH, USA  
Case Western Reserve University, USA  
Cancer Research UK London, UK  
Iowa State University, USA

DILS Steering Committee

Sarah Cohen-Boulakia  
Graham Kemp  
Ulf Leser  
Paolo Missier  
Norman Paton  
Louisa Raschid  
Erhard Rahm  

LRI, University of Paris-Sud 11, France  
Chalmers University of Technology, Sweden  
Humboldt-Universität zu Berlin, Germany  
Newcastle University, UK  
University of Manchester, UK  
University of Maryland, USA  
University of Leipzig, Germany

Organizing Committee

Greg Butler  
Christopher J.O. Baker  
Michel Dumontier  

Concordia University, Montreal, Canada  
University of New Brunswick, Saint John, Canada  
Carleton University, Ottawa, Canada

Webmaster

Artjom Klein  

University of New Brunswick, Saint John, Canada
# Table of Contents

Changing the Model in Pharma and Healthcare – Can We Afford to Wait Any Longer? ............................................. 1  
*Erich Alfred Gombocz*

Ibidas: Querying Flexible Data Structures to Explore Heterogeneous Bioinformatics Data ............................................. 23  
*Marc Hulsman, Jan J. Bot, Arjen P. de Vries, and Marcel J.T. Reinders*

From Questions to Effective Answers: On the Utility of Knowledge-Driven Querying Systems for Life Sciences Data .................................................................................................................. 38  
*Amir H. Asiaee, Prashant Doshi, Todd Minning, Satya Sahoo, Priti Parikh, Amit Sheth, and Rick L. Tarleton*

OmixAnalyzer – A Web-Based System for Management and Analysis of High-Throughput Omics Data Sets ............................................. 46  
*Thomas Stoltmann, Karin Zimmermann, André Koschmieder, and Ulf Leser*

The RDF Pipeline Framework: Automating Distributed, Dependency-Driven Data Pipelines ............................................. 54  
*David Booth*

Towards Interoperable BioNLP Semantic Web Services Using the SADI Framework ............................................. 69  
*Ahmad C. Bukhari, Artjom Klein, and Christopher J.O. Baker*

Optimizing Similarity Computations for Ontology Matching - Experiences from GOMMA ............................................. 81  
*Michael Hartung, Lars Kolb, Anika Groß, and Erhard Rahm*

Semi-automatic Adaptation of Mappings between Life Science Ontologies ............................................. 90  
*Anika Groß, Julio Cesar Dos Reis, Michael Hartung, Cédric Pruski, and Erhard Rahm*

Next Generation Cancer Data Discovery, Access, and Integration Using Prizms and Nanopublications ............................................. 105  
*James P. McCusker, Timothy Lebo, Michael Krauthammer, and Deborah L. McGuinness*