# Semantically Integrating Biomedical Databases to Support Inference

**Kevin M. Livingston, Michael Bada, William A. Baumgartner Jr., Lawrence E. Hunter**

Computational Bioscience Program, University of Colorado Anschutz Medical Campus, Aurora, CO, USA
{kevin.livngston, mike.bada, william.baumgartner, larry.hunter}@ucdenver.edu

## Motivation

Biomedical data integration currently exists as a "loose federation of bio-nations." (Goble 2008) While work is being done to integrate data sources it often involves "'semantic creep' – timid, piecemeal and ad hoc adoption of parts of standards." (Good 2006) Linking data across resources is necessary for building integrated systems (e.g. Belleau 2008, Ruttenberg 2009); however, linking the data without understanding the semantics of those links merely generates more data (Jain 2010). An integrated knowledge base is required for understanding biomedical systems by enabling queries over common biological representations.

## Methods

We modeled information about humans and seven major model organisms from 20 existing data sources using a common record representation. Forward-chaining rules are used to extract information implicit in the records and then explicitly reify the biomedical knowledge in OWL in terms of 14 prominent OBOs (Open Biomedical Ontologies). All data is stored in an RDF triplestore, and intermediate triple files are preserved. Queries are performed using SPARQL.

## Results and Discussion

The methods used in constructing the knowledge base provide several distinct advantages. By modeling the data sources in a common model new sources can be added quickly. This record data serves as provenance for the biomedical representations derived from it using declarative rules, and it remains available for querying alongside the biomedical representations.

When integrating multiple diverse data sets in this fashion, problems of identity and semantic mismatch arise. Mappings provided by the data sources are used to ensure identifiers from different sources that represent the same thing all point to the same biomedical entity. We make extensive use of subclassing in our modeling to help preserve consistency by avoiding conflicts and to provide monotonicity. This also helps alleviate issues of semantic mismatch (e.g., combining species-level and strain-level data sources). Our approach also allows us to simultaneously model multiple common abstractions that might be inconsistent with each other. For example, while genes, RNA, and proteins are all distinct "things" it is common to refer to gene-or-gene-product abstractions in biology. We explicitly represent these abstractions and make them available for modeling and querying.

We have built KaBOB (Knowledge Base of Biomedicine) by integrating information from over 20 existing biomedical data sources about humans and seven major model organisms. The knowledge base is modeled in OWL and grounded in the OBOs. It is comprised of over 419 million RDF triples. Queries can be posed to KaBOB in terms of biomedical concepts and abstractions, instead of requiring knowledge of source-specific encodings and terminology.



## References

Belleau, F., Nolin, M. A., Tourigny, N., Rigault, P., & Morissette, J. (2008). Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform , 41(5), 706-716.

Goble, C., & Stevens, R. (2008). State of the nation in data integration for bioinformatics. J Biomed Inform , 41(5), 687-693.

Good, B. M., & Wilkinson, M. D. (2006). The life sciences semantic web is full of creeps!. Brief Bioinform, 7(3), 275-286.

Hitzler, P. (2009). Towards reasoning pragmatics. In GeoSpatial Semantics (pp. 9-25). Springer Berlin Heidelberg.

Jain, P., Hitzler, P., Yeh, P. Z., Verma, K., & Sheth, A. P. (2010). Linked Data Is Merely More Data. In AAAI Spring Symposium: linked data meets artificial intelligence.

Ruttenberg, A., Rees, J. A., Samwald, M., & Marshall, M. S. (2009). Life sciences on the Semantic Web: the Neurocommons and beyond. Brief Bioinform, 10(2), 193-204.