Evaluation of ontology-powered scientific research as a means to assess and improve ontology quality

Michel Dumontier, **Ph.D.**

Associate Professor of Bioinformatics

Department of Biology, School of Computer Science, Institute of Biochemistry, Carleton University

Ottawa Institute of Systems Biology

Ottawa-Carleton Institute of Biomedical Engineering

Professeur Associé, Université Laval

Chair, W3C Semantic Web for Health Care and Life Sciences Interest Group

Why should users care about what terms an ontology contains and how it is structured?

How should ontology designers evaluate their research?

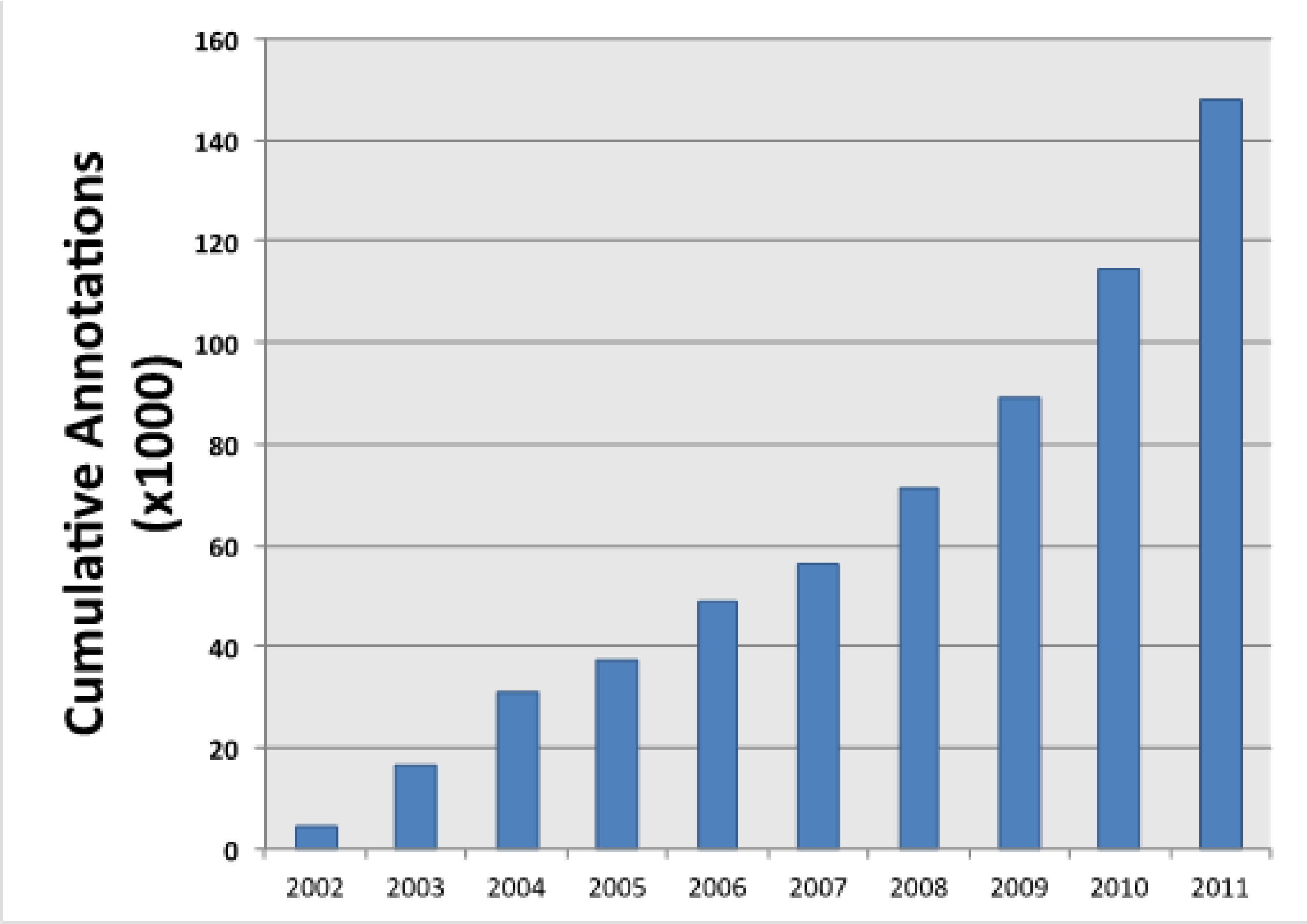
Use of ontologies in biomedical investigations

- In 1998, researchers involved in annotating fruit fly, mouse and yeast genomes came together to build the Gene Ontology (GO) - a controlled vocabulary to annotate genes (gene products) with
  - Molecular function
  - Cellular compartment
  - Biological process
- Back in 2006, the cost of developing the GO was estimated to be >$16M
- Thousands of genomes have been annotated with nearly 30,000 terms.
- Hundreds of tools have been devised to mine this information in order to help elucidate organismal capability and limitations, and to interpret the results of experiments

Gene Set Enrichment Analysis

- **Goal**: identify a set of terms that are significantly enriched for a set of genes identified through some experiment
- Compare the set of annotations for target genes against all other plausible genes (Fisher's exact test).
- Depends on
  - # and structure of terms in the ontology
  - # of annotations using ontology terms

Continuous growth in gene ontology annotations



What’s the impact of changes in the gene ontology and annotations on gene set enrichment analysis?

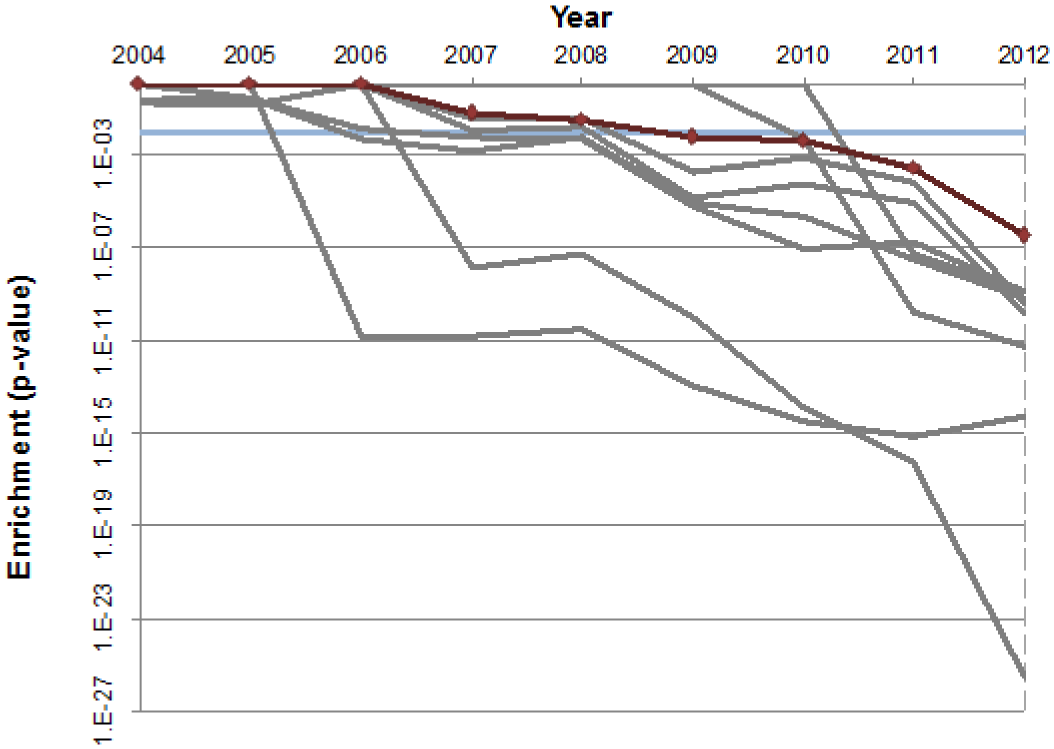
Erik L. Clarke, Benjamin M. Good, and Andrew I. Su. A Task-Based Approach For Gene Ontology Evaluation. Bio-ontologies 2012 SIG.

Top 10 most enriched terms *differ* in subsequent years

2006	**2012**
System development	Synaptic transmission
Cell-cell signaling	System development
Cell communication	Response to interferon-γ
Microtubule-based process	Secretion by cell
Nervous system development	Secretion
Inositol lipid-mediated signaling	Chemotaxis
Phosphatidylinositol-mediated signaling	Taxis
Regulation of catalytic activity	Blood coagulation
Regulation of cell cycle	Coagulation
Intracellular protein transport	Cellular response to interferon-γ

Erik L. Clarke, Benjamin M. Good, and Andrew I. Su. A Task-Based Approach For Gene Ontology Evaluation. Bio-ontologies 2012 SIG.

Significance of any given term changes with time



P-Values of Angiogenesis (red) and Ten Top Terms (grey) in 2012 for GDS1962
The blue line is the significance threshold (p-value < 0.01).

Angiogenesis only becomes significant after 2007.

Eight terms only become significant after 2006.

Conclude: enrichment analysis using human Gene Ontology annotations improved significantly since 2002

Erik L. Clarke, Benjamin M. Good, and Andrew I. Su. A Task-Based Approach For Gene Ontology Evaluation. Bio-ontologies 2012 SIG.

A Task-Based Approach For Gene Ontology Evaluation

- Ontology-based research is not future proof.
- Re-analysis of past experiments may yield new and important results. However, it may also remove previously significant results
- Suggests that continuous evaluation of research results needs to occur.

- We need to understand how changes in ontologies affect our research results

Evaluation of Ontology Research

- Considerable debate about the importance and effectiveness of metrics to evaluate results of ontology research
- What constitutes a (novel) research result?
  - Capability to do X via some method
  - Improved capability to do X, assessed by methodological comparison
- Challenges in ontology design
  - Coverage of domain and degree of formalization are limiting factors
  - A combination of factors are likely required to predict the capability of an ontology for an arbitrary scenario.

Hoehndorf R, Dumontier M, Gkoutos GV. Evaluation of research in biomedical ontologies.. Brief Bioinform. 2012

Quantifying Ontology Research

Application	Evaluation	Description
Community agreement	User-study [% agreement, κ statistic]	From textual descriptions to any aspect of formalization, generate confidence measures that indicate the degree to which a significant number [>15] of people agree.
Consistent data annotation	User-study [% agreement, κ statistic]	Use an ontology to annotate the types, attributes and relations in a dataset
Data integration	Analysis [precision, recall, F-measure]	Establish agreement on the points of integration and/or provide an analysis of integrated data set, compare to use cases or gold standard.
Query answering	Test suite [# of tests passed, precision, recall, f-measure, complexity class]	Evaluate the extent to which the ontology can be used to answer questions of relevance to the domain. Use or jointly establish a gold standard with other communities.
Data consistency	Test suite [# of tests passed, contradictions found, complexity class]	Evaluate the extent to which the ontology can be used to identify inconsistent knowledge.
Novel scientific results	Case-specific validation [p-value, f-measure, ROC AUC]	Evaluate the extent to which novel relations can be extracted against some gold standard.

Quantifying Ontology Research

- **Community agreement**
  - Assess the degree to which a community agrees about any aspect of an ontology, for example:
    - Evaluate alternate textual definitions,
    - Associate and evaluate synonyms, hyponyms
    - Associate and evaluate mereological, subsumption and other relations
  - Quantitatively asses with user-study [% agreement, κ statistic]
  - Example: 39% chance that GO curators select the same GO term to annotate text; 19% chance they will annotate a term from the same GO lineage and 43% chance to extract a term from a new/different lineage. [1]

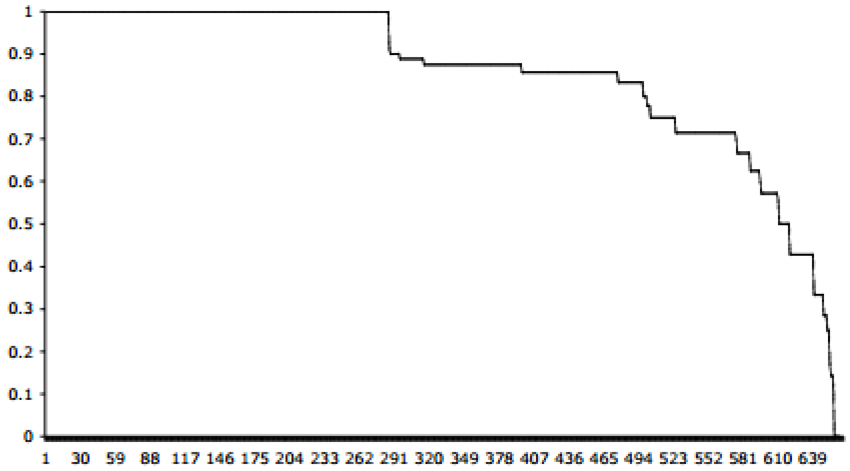
[1] **Evaluation of GO-based functional similarity measures using *S. cerevisiae* protein interaction and expression profile data.** *BMC Bioinformatics* 2008, **9**:472 doi:10.1186/1471-2105-9-472

Fast, Cheap and Out of Control: A Zero Curation Model for Ontology Development

Benjamin M. Good, Erin M. Tranfield, Poh C. Tan, Marlene Shehata, Gurpreet K. Singhera, John Gosselink, Elena B. Okon, and Mark D. Wilkinson

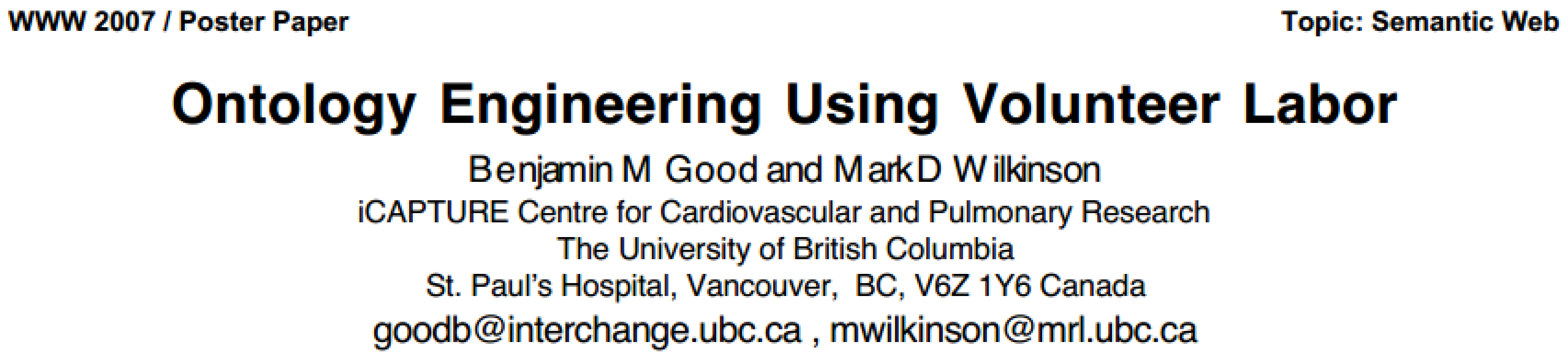
Pacific Symposium on Biocomputing 11:128-139(2006)

- 68 volunteers linked 661 terms to each other and to a pre-existing upper ontology by adding 245 hyponym relationships and 340 synonym relationships
  - Judged terms to be sensible, nonsense, or outside their expertise



Less than 50% of terms had 100% agreement. Another 30% had 70-90% agreement.

Would you include the remaining 20% in your ontology?



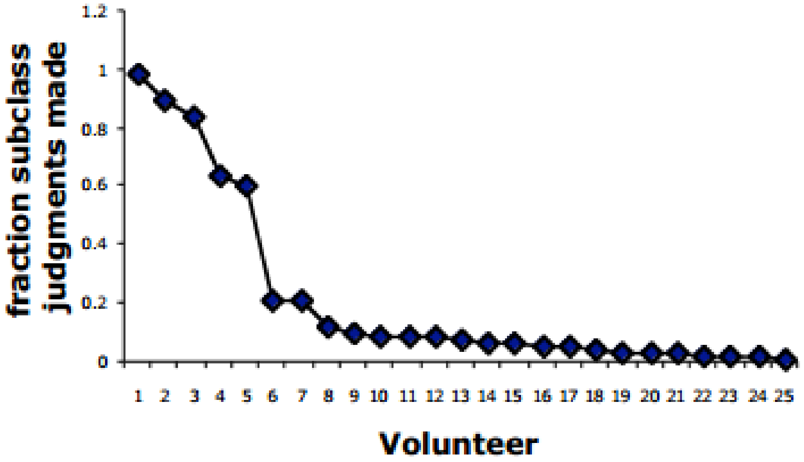
WWW 2007 / Poster Paper

Topic: Semantic Web

Ontology Engineering Using Volunteer Labor

Benjamin M Good and Mark D Wilkinson
iCAPTURE Centre for Cardiovascular and Pulmonary Research
The University of British Columbia
St. Paul's Hospital, Vancouver, BC, V6Z 1Y6 Canada
goodb@interchange.ubc.ca , mwilkinson@mrl.ubc.ca

Used volunteers to judge the correctness of automatically inferred subsumption relationships, generated from an automatic mapping of MeSH to OWL (expect ~40% incorrect subclass relations)
- 130 subclass relations tested with 25 volunteers



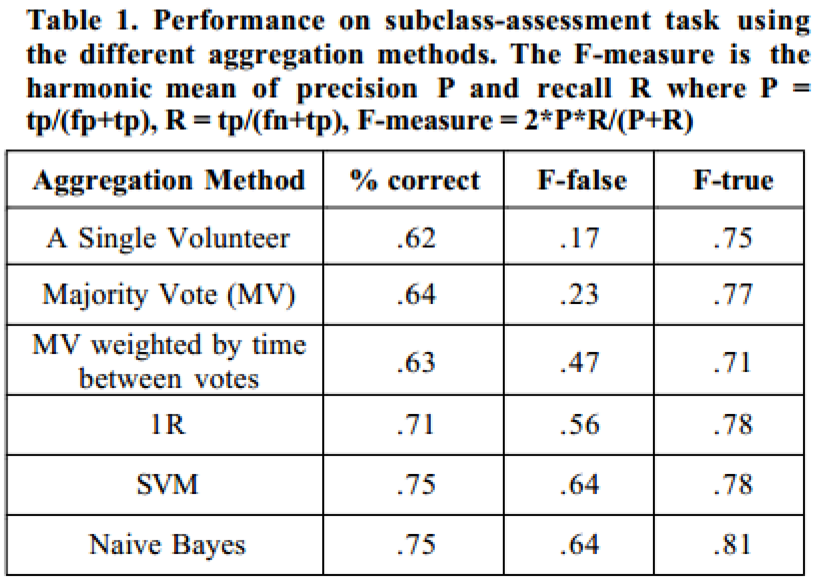


Table 1. Performance on subclass-assessment task using the different aggregation methods. The F-measure is the harmonic mean of precision P and recall R where P = tp/(fp+tp), R = tp/(fn+tp), F-measure = 2*P*R/(P+R)

Aggregation Method	% correct	F-false	F-true
A Single Volunteer	.62	.17	.75
Majority Vote (MV)	.64	.23	.77
MV weighted by time between votes	.63	.47	.71
1R	.71	.56	.78
SVM	.75	.64	.78
Naive Bayes	.75	.64	.81

confidence weighted response

Ontology-based
Data Integration, Consistency Checking and Discovery

- **Checking the consistency of semantic annotations [1]**
  - Formalized semantic annotations in SBML models as OWL axioms. Automated reasoning uncovered inconsistencies in 16 models.
    - e.g. alpha-D-glucose phosphate is not the required ATP in an ATP-dependent reaction (required GO + ChEBI + disjoint + existential + universal quantification)
- **Finding significant biomedical associations [2]**
  - found significant associations between genes, drugs, diseases and pathways using Drugbank, PharmGKB, CTD, PID across *categories* of drugs (ChEBI, ATC, MeSH) and diseases (DO, MeSH)
  - 22,653 pathway-disease type associations (6304 over; 16,349 under)
    - carcinosarcoma (DOID:4236) and Zidovudine Pathway (PharmGKB:PA165859361)
  - 13,826 pathway-chemical type associations (12,564 over; 1262 under)
    - drug clopidogrel (CHEBI:37941) with Endothelin signaling pathway (PharmGKB:PA164728163);

http://pharmgkb-owl.googlecode.com

1. Integrating systems biology models and biomedical ontologies. BMC Systems Biology. 2011. 5 : 124
2. Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics. Bioinformatics. 2012. *in press*

HyQue

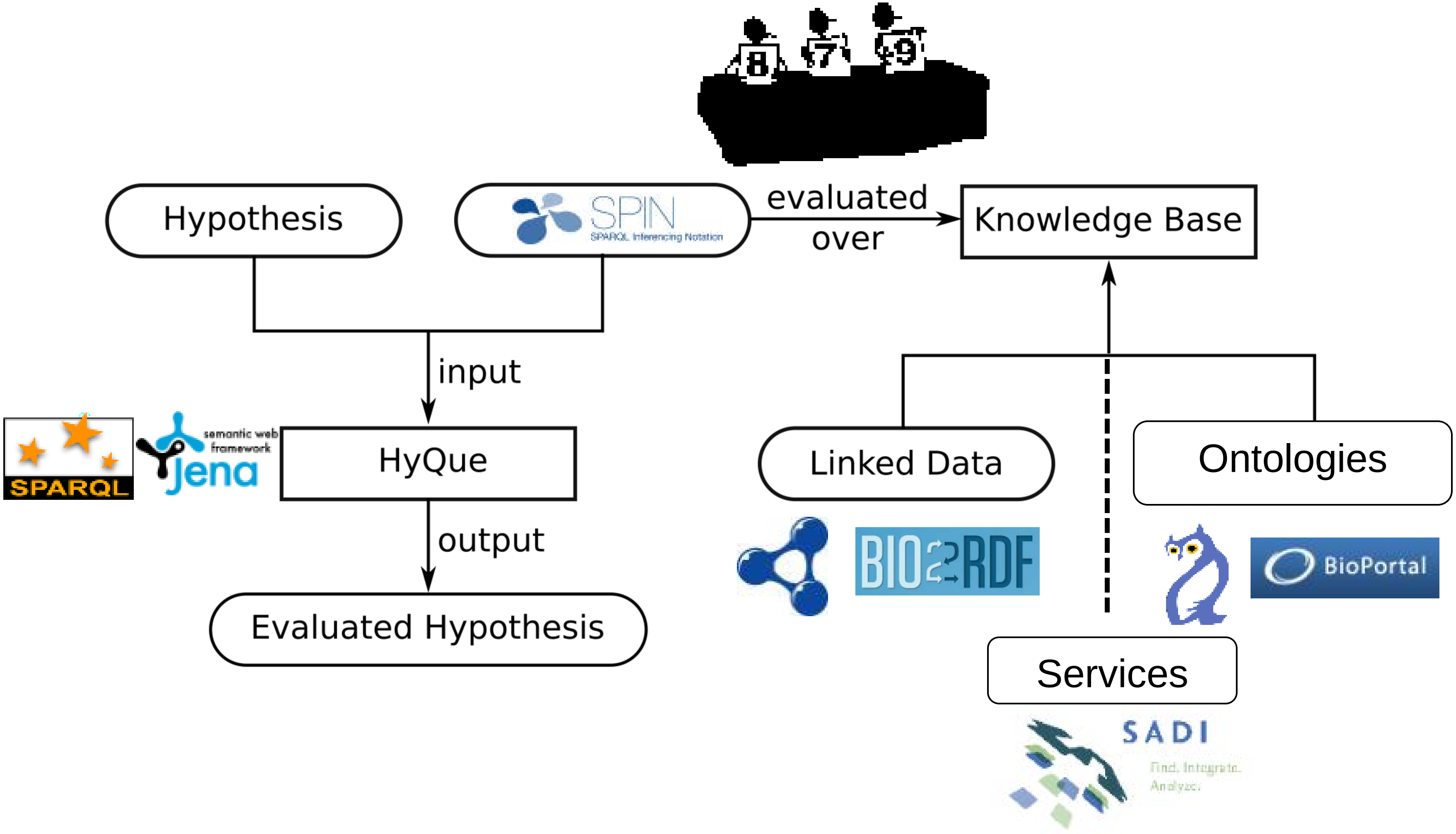
HyQue is the Hypothesis query and evaluation system

- A platform for knowledge discovery
- Facilitates hypothesis formulation and evaluation
- Leverages Semantic Web technologies to provide access to facts, expert knowledge and web services
- Conforms to a simplified event-based model
- Supports evaluation against positive and negative findings
- Transparent and reproducible evidence prioritization
- Provenance of across all elements of hypothesis testing
  - trace a hypothesis to its evaluation, including the data and rules used

Evaluating scientific hypotheses using the SPARQL Inferencing Notation. Extended Semantic Web Conference (ESWC 2012). Heraklion, Crete. May 27-31, 2012.

HyQue: evaluating hypotheses using Semantic Web technologies. J Biomed Semantics. 2011 May 17;2 Suppl 2:S3.

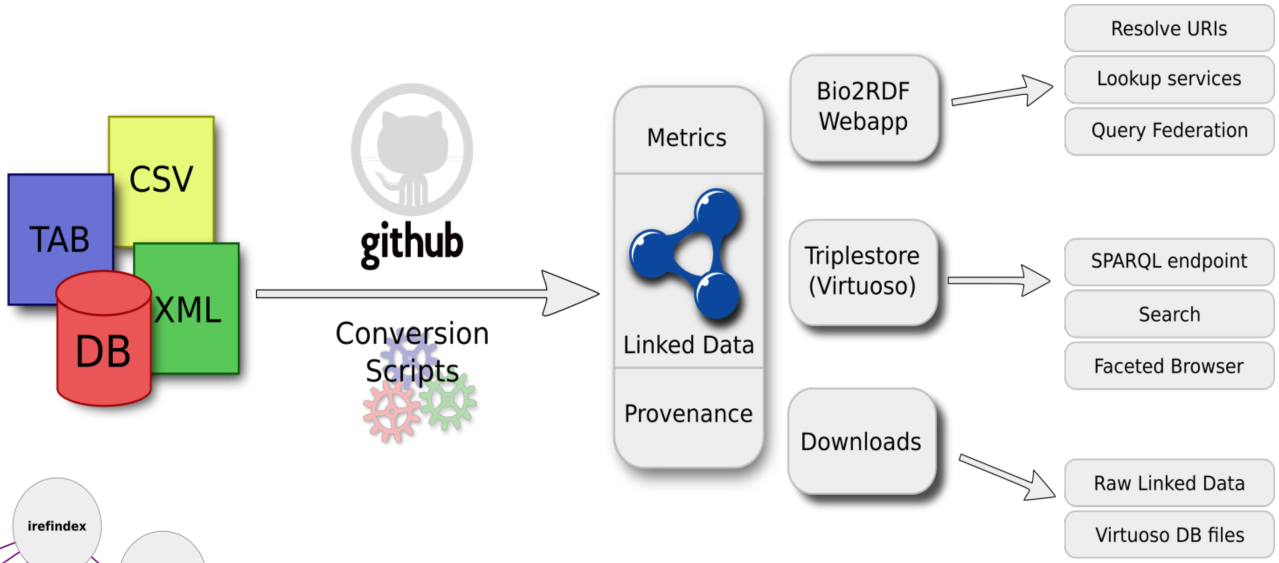
HyQue Architecture



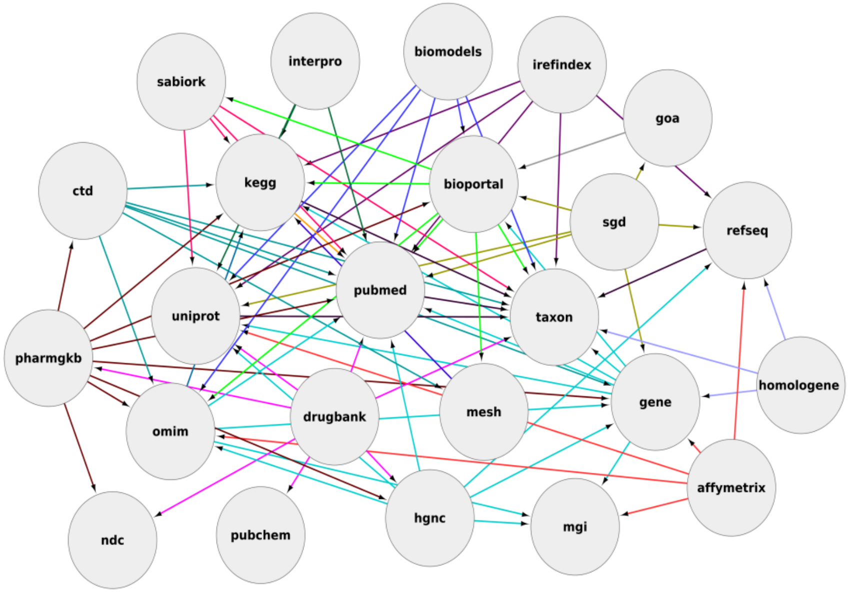
BIO2RDF

At the heart of Linked Data for the Life Sciences

chemicals/drugs/formulations,
genomes/genes/proteins, domains
Interactions, complexes & pathways
animal models and phenotypes
Disease, genetic markers, treatments
Terminologies & publications

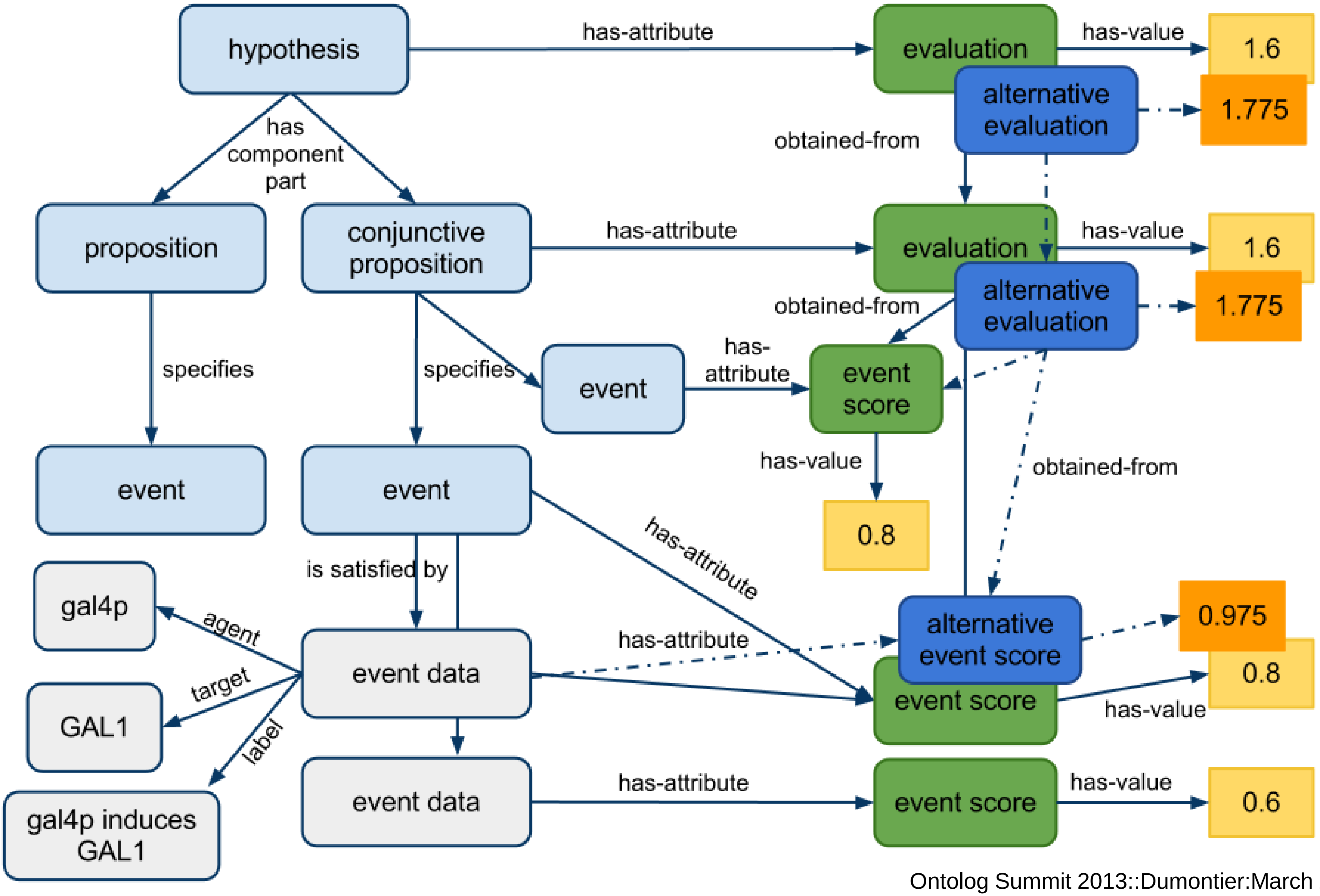


CSV
TAB
XML
DB
github
Conversion Scripts
Metrics
Linked Data
Provenance
Bio2RDF Webapp
Triplestore (Virtuoso)
Downloads
Resolve URIs
Lookup services
Query Federation
SPARQL endpoint
Search
Faceted Browser
Raw Linked Data
Virtuoso DB files



- Free and open source
- Based on Semantic Web standards
- Billions of interlinked statements from dozens of conventional and high value datasets
- Partnerships with EBI, NCBI, DBCLS, NCBO, OpenPHACTS, and commercial tool providers

Customization of rules and rulesets may lead to different evidence-based evaluations



Summary

- Quantitative comparison and evaluation is at the heart of the scientific enterprise.
- Scientists that make use of ontologies should control for and quantitatively assess the contribution of any ontology component.
- Ontology designers must include quantitative evaluation to sustain any claims about community agreement, semantic annotation, consistency checking, query answering, or enabling new scientific results.
- We can build on knowledge sharing platforms like Bio2RDF and hypothesis testing platforms like HyQue to undertake and evaluate ontology-based research.

dumontierlab.com

michel_dumontier@carleton.ca

Website: [*http://dumontierlab.com*](http://dumontierlab.com)
Presentations: http://slideshare.com/micheldumontier