

2009



ANITA BORG INSTITUTE  
FOR WOMEN AND TECHNOLOGY



Association for  
Computing Machinery

# Towards the Semi-Automated Building of Knowledge Bases for Biological Research

Natalia Villanueva-Rosales  
Supervised by: Michel Dumontier  
Carleton University, CA  
PhD Forum 3

# Current Web vs. Semantic Web (SW)



← Natalia  
www.natalia-villanueva.com

Mexicans love Mexican food  
(Mexican culture book)

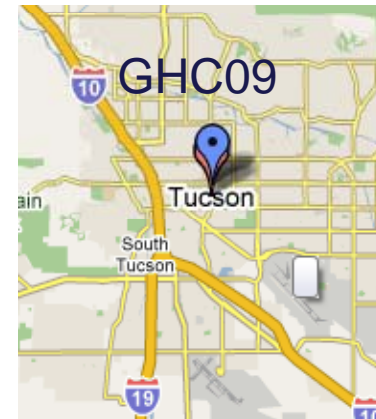
Q. Mexican food lovers currently in Tucson

- Current syntactic web ☹, ?
- Semantic Web (SW) 😊, → Natalia
  - Machine understandable www



Natalia is attending GHC09

RSS, Natalia es mexicana



GHC09: Sep. 30 - Oct 3, 2009



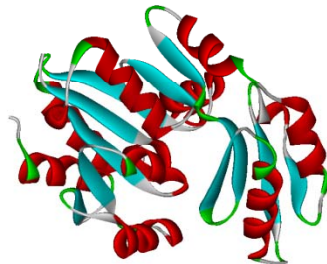
THE GRACE HOPPER CELEBRATION  
OF WOMEN IN COMPUTING

2009

# Semantic Web (SW) for Biological Research

## Me

- Homo sapiens, tax\_id:9606
- ~ 25,000 genes, e.g. ABCB1,
- ~ 33,000 proteins, e.g. NP\_596892.1
  - Proteins composed by amino acids, e.g. Sucrose-phosphatase
- Genetic background (Mexican)
  - Predisposition
  - ABCB1\_3435\_C



## Drugs

- Treatment for disease
- Effects may change depending on genetic backgrounds



What is the best treatment for Natalia for depression?

- Personalized medicine
- Manual integration/curation/querying by experts. [BIB08]
- NortriptylineABCB1Treatment1



THE GRACE HOPPER CELEBRATION  
OF WOMEN IN COMPUTING

2009

# Problem Statement

## Semantic Web (SW)

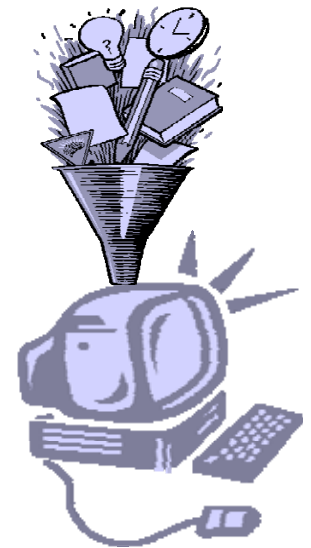
- Machine understandable knowledge for automated reasoning.
- Ontologies manually created/curated (**bottleneck**), DB vs. KB.

## But...

- Structured knowledge is captured in DB.
- Agreement between designers and users (domain).

## A possible solution...

- Extract knowledge (not only data) from databases, with more natural representations.
- **New!** W3C working group RDB2RDF, new standard.



# Related work

- **Ontology population tools** [D2RServer, DataMaster, ...]
  - Simple mappings, encoded, manual creation/refinement.
  - Mixed model semantics: Database model, not only domain knowledge
  - RDF, Frames or OWL 1.0.
- **Ontology extraction methodologies** [Man et al. '05, Shen et al. '06, DB2OWL, ...]
  - Mixed model semantics
  - Limited expressivity.
  - Some semantics loss.
- **Reverse Engineering!** [Navathe & Awong '88, Chiang et al. '94]
  - Extract ER model
  - Assumptions: DB designed with standard methodology and practices (normalization), availability of all keys, particularly primary keys.



ANITA BORG INSTITUTE  
FOR WOMEN AND TECHNOLOGY



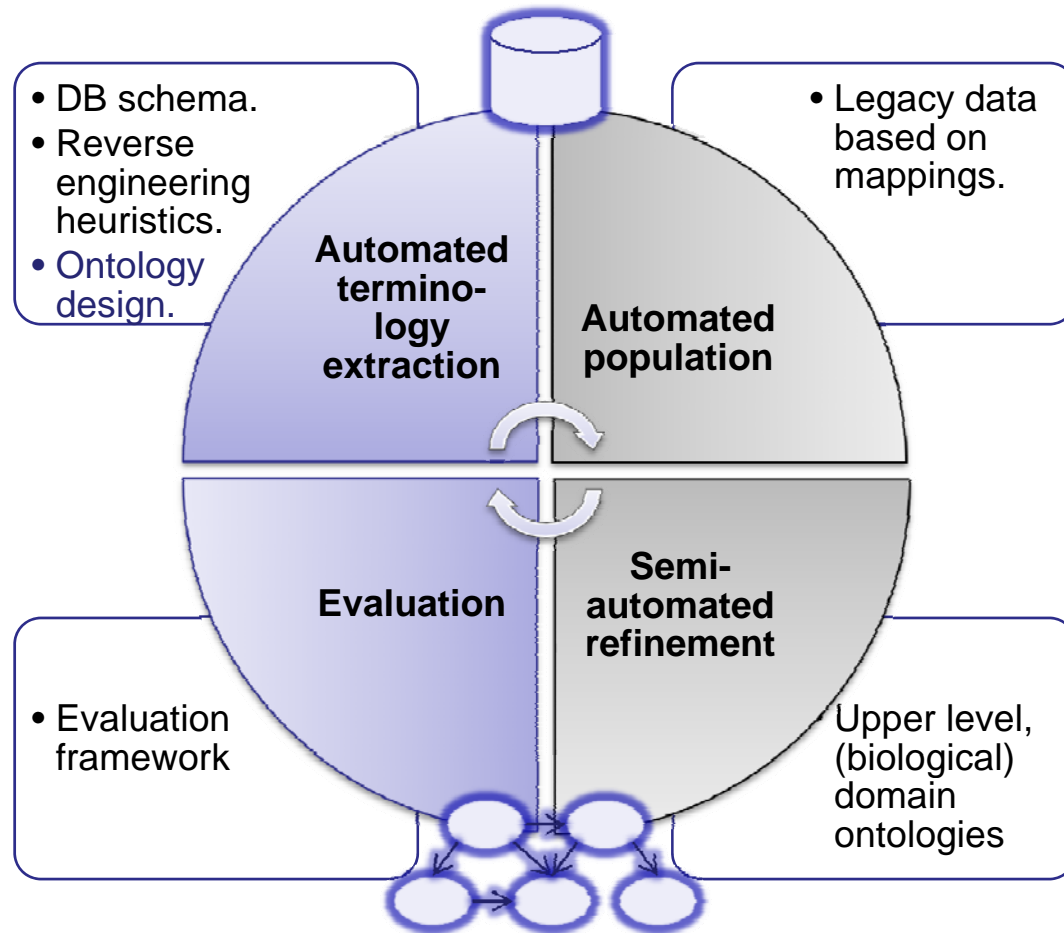
Association for  
Computing Machinery

THE GRACE HOPPER CELEBRATION  
OF WOMEN IN COMPUTING

2009

# This Research

## Ontology Extraction from DB



In Progress /  
Done

To do



THE GRACE HOPPER CELEBRATION  
OF WOMEN IN COMPUTING

2009

# Evaluation framework



- OWL Ontologies with SWRL rules
  - Relational model OWL ontology
    - App. access to DB schema as an instance
    - Classification of relations according to the keys and attributes they contain.
  - Relational model to OWL mapping OWL ontology + SWRL rules to capture mappings.
    - Explicit mappings
    - Machine understandable for automated reasoning
  - Comparison of related work available as ontologies.
    - Automated creation and population of OWL ontologies.
- Ontology evaluation.
- Testing sets: SGD/yOWL, PharmGKB/Pharmacogenomics ontology.



ANITA BORG INSTITUTE  
FOR WOMEN AND TECHNOLOGY

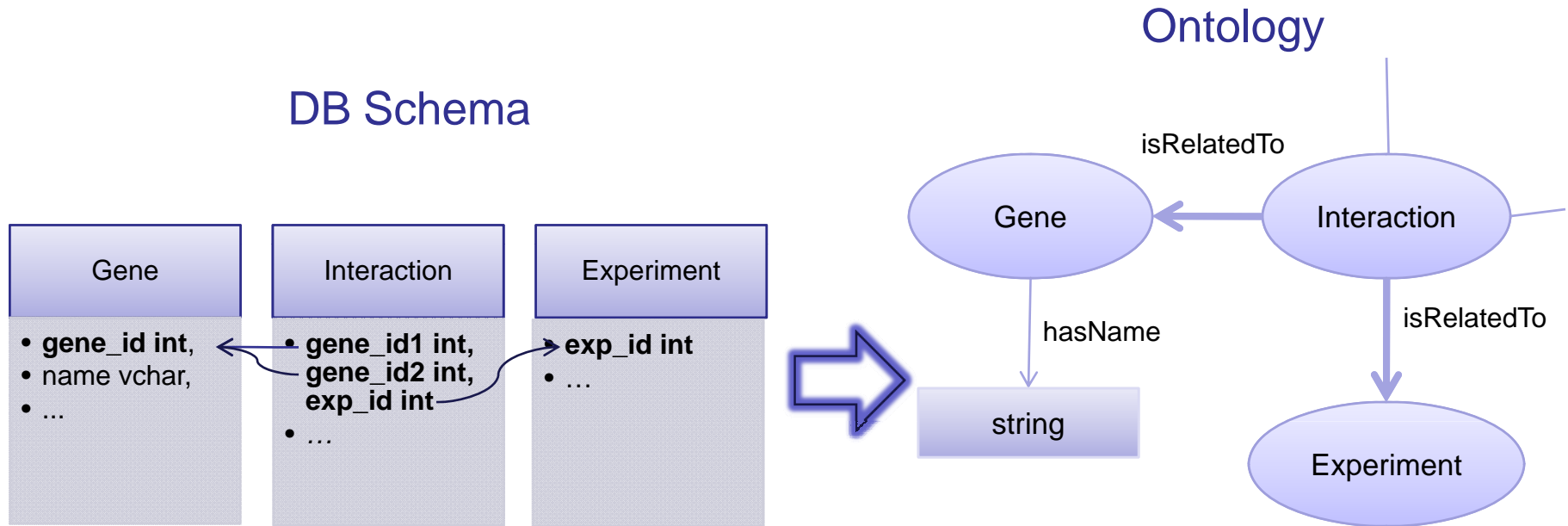


Association for  
Computing Machinery

THE GRACE HOPPER CELEBRATION  
OF WOMEN IN COMPUTING

2009

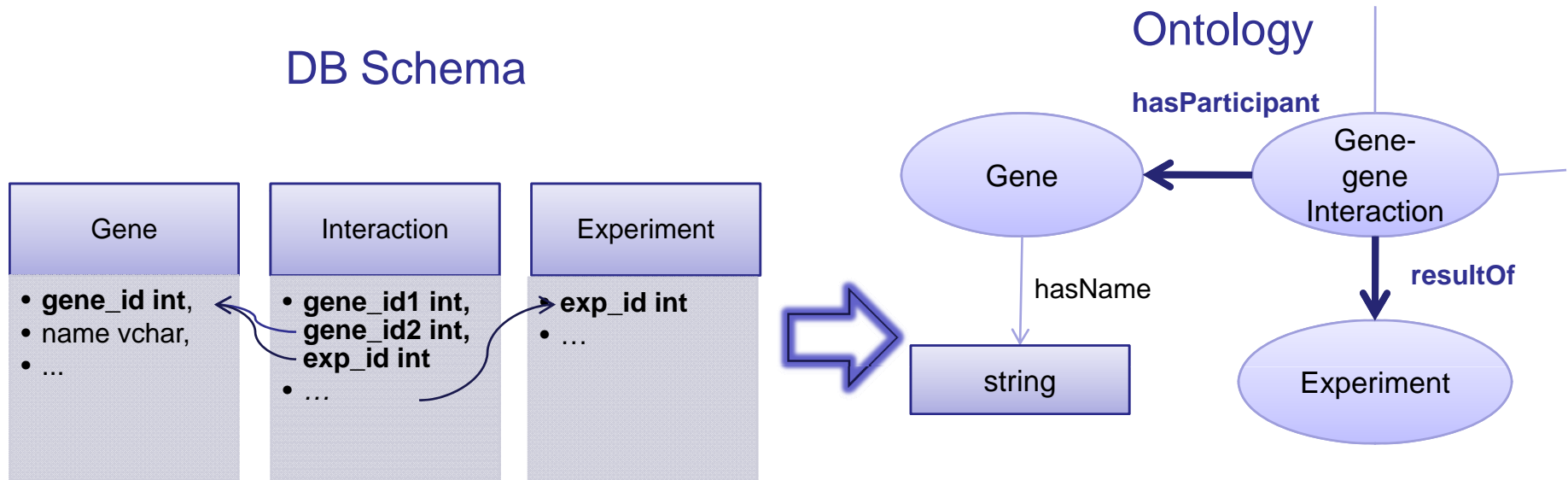
# N-ary relation extraction



Interaction table: Attributes of the primary key are foreign keys to primary keys in other tables.

➔ Necessary link.

# N-ary relation refinement



Domain Knowledge: Gene is an Object, Interaction and Experiment are Processes

# So far

- Initial Heuristics
  - Concept extraction
    - Classes, hierarchical class relations.
  - Property extraction
    - N-ary (binary) object properties, datatype properties.
  - Constraints axioms (consistency)
    - Disambiguation w/unique identifiers, functional dependencies, cardinality constraints, existential, universal constraints.
  - Separation of model constraints from domain constraints.
- Implementation in JRuby/Java: OWL API, Pellet/FaCT++
  - Interoperability, standardization.



ANITA BORG INSTITUTE  
FOR WOMEN AND TECHNOLOGY



Association for  
Computing Machinery

THE GRACE HOPPER CELEBRATION  
OF WOMEN IN COMPUTING

2009

# Summary

- Semantic Web for Biological Research
  - **Promising** platform.
  - Integration, exposure and **query answering** of data knowledge.
  - Knowledge acquisition **bottleneck**. (DB → RDF, OWL)
- This work
  - **Improve** previous approaches
    - Higher level of **automation**
    - More **expressive knowledge**
    - Model semantics vs. domain semantics..
  - Provide an evaluation framework
  - Realize the Semantic Web!



THE GRACE HOPPER CELEBRATION  
OF WOMEN IN COMPUTING

2009

THE GRACE HOPPER CELEBRATION OF WOMEN IN COMPUTING

2009



ANITA BORG INSTITUTE  
FOR WOMEN AND TECHNOLOGY



Association for  
Computing Machinery

Thanks!  
Questions?

[www.natalia-villanueva.com](http://www.natalia-villanueva.com)

[www.dumontierlab.com](http://www.dumontierlab.com)