

Integrating Hierarchical Controlled Vocabularies with OWL Ontology

Melissa J. Davis

ARC Centre of Excellence in Bioinformatics

and

Institute for Molecular Bioscience, University of Queensland



THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA

IMB *Institute for Molecular Bioscience*






BioMANTA

Modelling
and
Analysis
of
Biological
Network
Activity

- Unravelling complex interactions between biological networks at varying levels of resolution requires systems approach
- Computational modelling and analysis of large-scale protein-protein interaction (PPI) and compound activity networks
- Semantic Web (SW) offers flexible technologies for creating meaningful representations of data on the web → *Semantic Interactome Model*
- Knowledge representation using SW standards Resource Description Framework (RDF) and Web Ontology Language (OWL) enables machine inference and facilitates knowledge discovery

Semantic Web

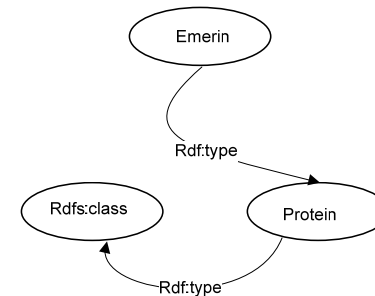
- **Current web:** Most information is natural language
 - Meaningful to human users who understand the meaning of natural language
- **Semantic Web:** standards for publishing machine-readable information on the web
 - Standard formats for integration and exchange of data (RDF)
 - Formal language to express semantics (the meaning of concepts) - OWL
 - Unambiguous representation

VPS9	Vacuolar sorting protein
Entry: YML097c	
Alias: VPL31; VPT9	
Classification: known protein 5423 Entries Evi PUBMED	
Feature Type: CDS	
Features	
 PROTEIN VIEW PEDANT help BLASTP PROSITE BLOCKS PFAM	
Similarity:	<p>Paralogs (14.4 %); Homologs in Hemiascomycota (88.7 %); Ascomycota (88.7 %); Fungi (88.7 %); Eukaryota (88.7 %); Plants (22.5 %); Mammalia (22.4 %); Human (20.8 %); Bacteria (15.1 %); All except yeast (88.7 %)</p> 
	<p>SESAM: Seed Extraction Sequence Analysis Method - 'Seed Extraction Sequence Analysis Method' to find Paralogs and Fungal Orthologs</p> <ul style="list-style-type: none"> ◊ similarity to human Ras inhibitor
Functional Classification:	<ul style="list-style-type: none"> ◊ CELLULAR TRANSPORT, TRANSPORT FACILITIES AND TRANSPORT ROUTES ..transport routes ...vacuolar lysosomal transport 154 Entries Evi PUBMED ◊ CELLULAR TRANSPORT, TRANSPORT FACILITIES AND TRANSPORT ROUTES ..transport routes ...vesicular transport (Golgi network, etc.) 200 Entries Evi ◊ PROTEIN FATE (folding, modification, destination) ..protein targeting, sorting and translocation 280 Entries Evi PUBMED ◊ REGULATION OF METABOLISM AND PROTEIN FUNCTION ..regulation of protein activity ...guanyl-nucleotide exchange factor (GEF) 18 Entries Evi PUBMED
InterPro:	<ul style="list-style-type: none"> ◊ IPR001005 Myb DNA-binding domain (Match details) 31 Entries ◊ IPR003123 Vacuolar sorting protein 9 (Match details) 2 Entries ◊ IPR003892 Ubiquitin system component Cue (Match details) 7 Entries
Localization:	<p>VPS9 localization details</p> <ul style="list-style-type: none"> ◊ cytoplasm
Protein Interactions and Complexes:	 Details of Interactions and Complexes on VPS9
Remarks:	<ul style="list-style-type: none"> ◊ residues 130-143 are predicted to form a coiled-coil domain ◊ residues 331-340 contain a highly charged patch of 10 contiguous aspartate and lysine residues

Semantic Web Technologies

- Resource Description Framework (RDF) and RDF Schema (RDFS)
 - RDF for resources; RDFS for vocabulary
 - Data presented as triples, graph or xml
<subject> <predicate> <object>

<Protein> <rdf:type> <rdfs:Class>
<Emerin> <rdf:type> <Protein>



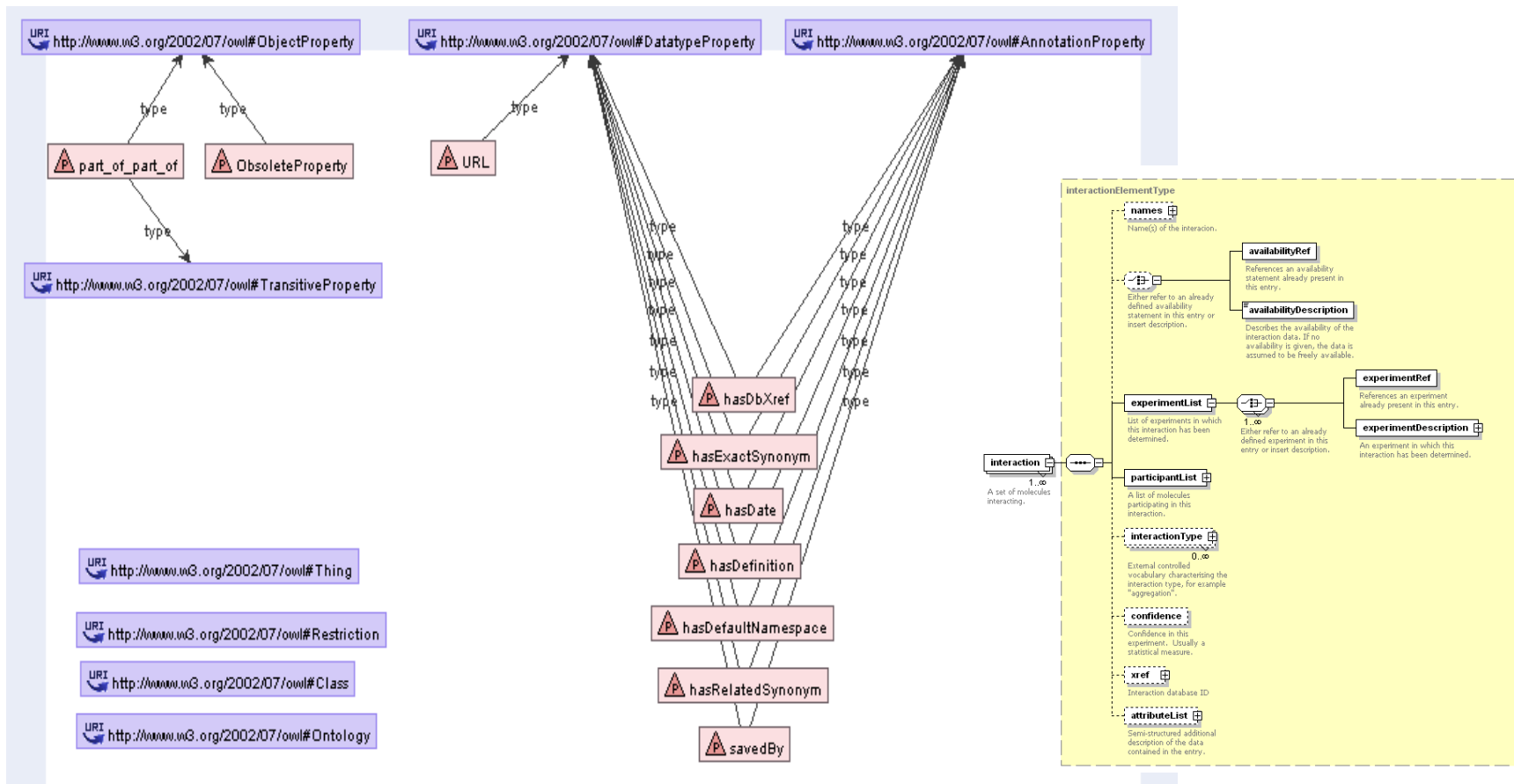
- OWL
 - Web Ontology Language for description of ontology
 - OWL is RDF: can be expressed in triples, graph or xml
 - cf Open Biomedical Ontology (OBO) format – familiar to users of GO
- Inference engines
 - software to reason about information in a knowledge representation and infer new data from what is presented
- SPARQL
 - query language for RDF
 - Returns results sets or RDF graphs

Knowledge representation with Ontology

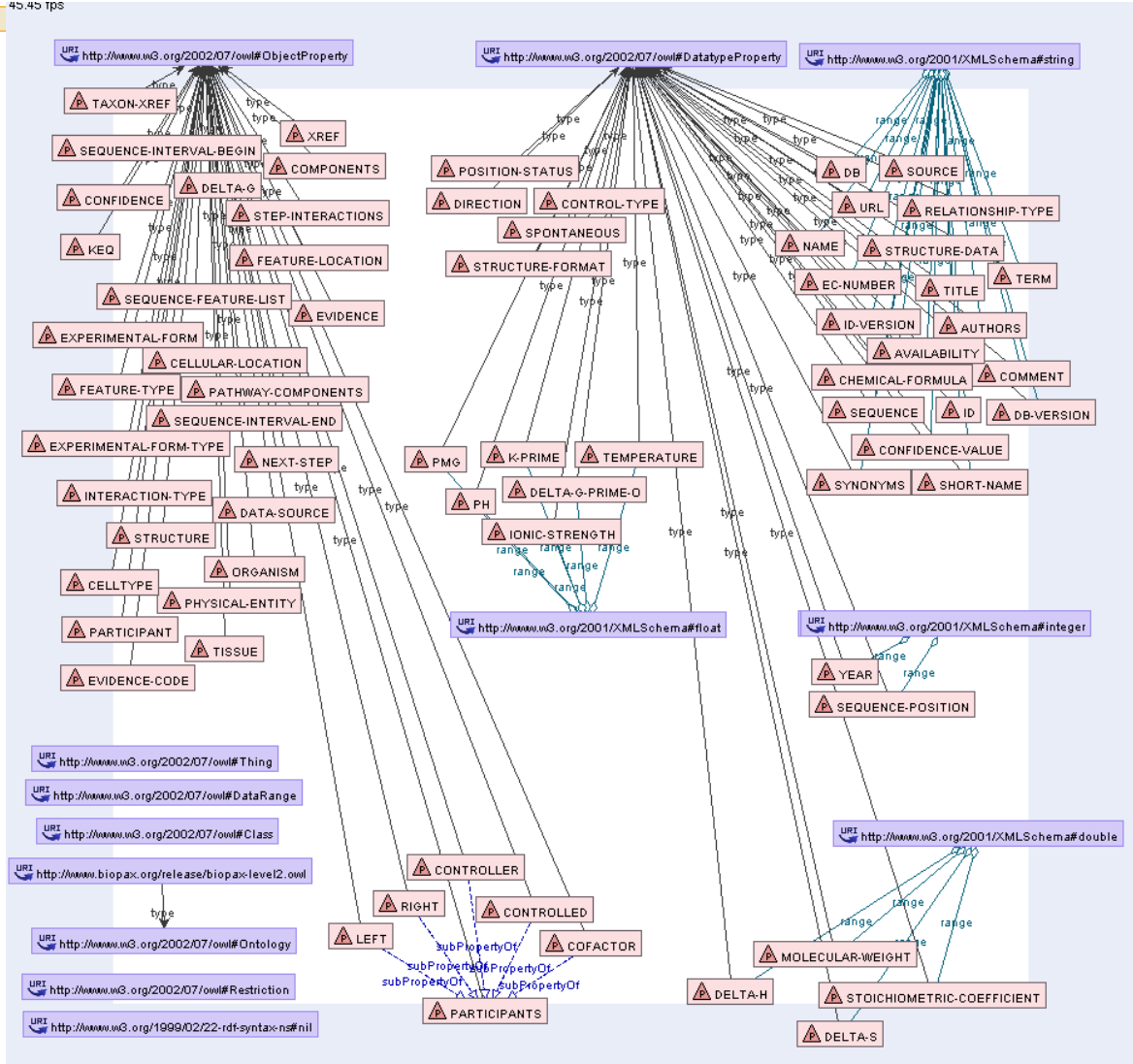
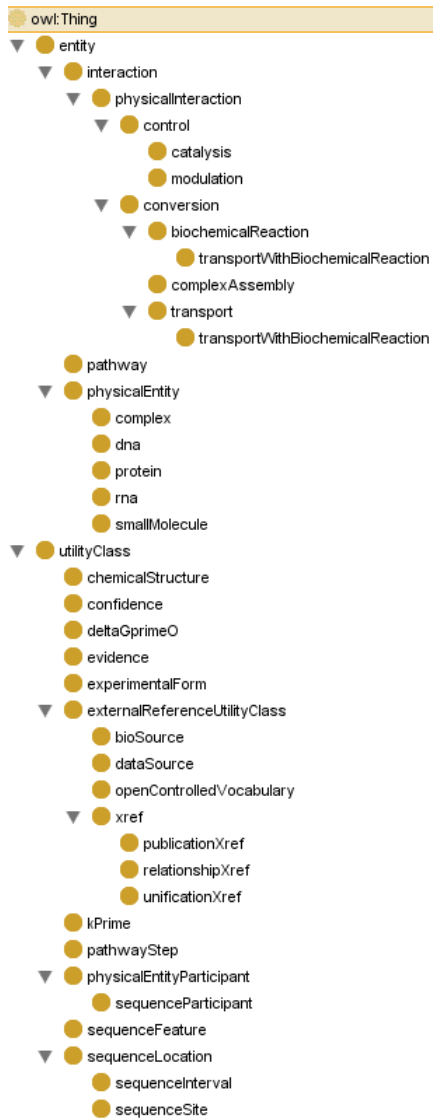
- Ontology can mean different things to different people, and frequently the intended use of the ontology shapes its structure and capability
 - Unstructured but controlled vocabulary
 - Structured, frequently hierarchical controlled vocabulary (eg: OBO)
 - Semantic web standard Web Ontology Language (OWL)
 - Lists of terms, hierarchies of terms, logically consistent knowledge representations from which new data may be inferred
 - Different approaches have strengths and weaknesses, well illustrated by two examples...

Current representations – PSI-MI...

- Protein Standards Initiative – Molecular Interactions (PSI-MI)
 - OBO format ontology
 - Hierarchical arrangement with deep coverage but limited relations



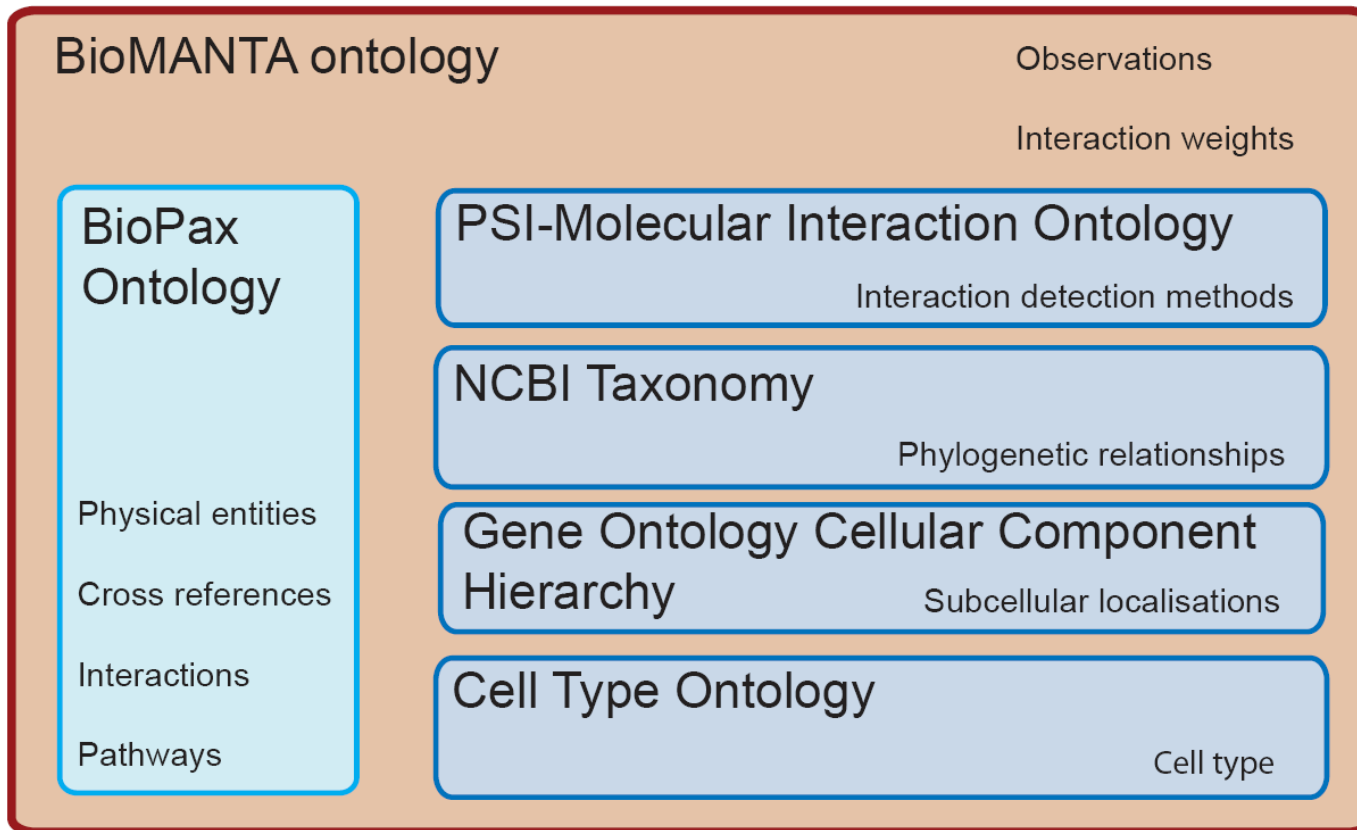
... and BioPAX



Ontology integration problem

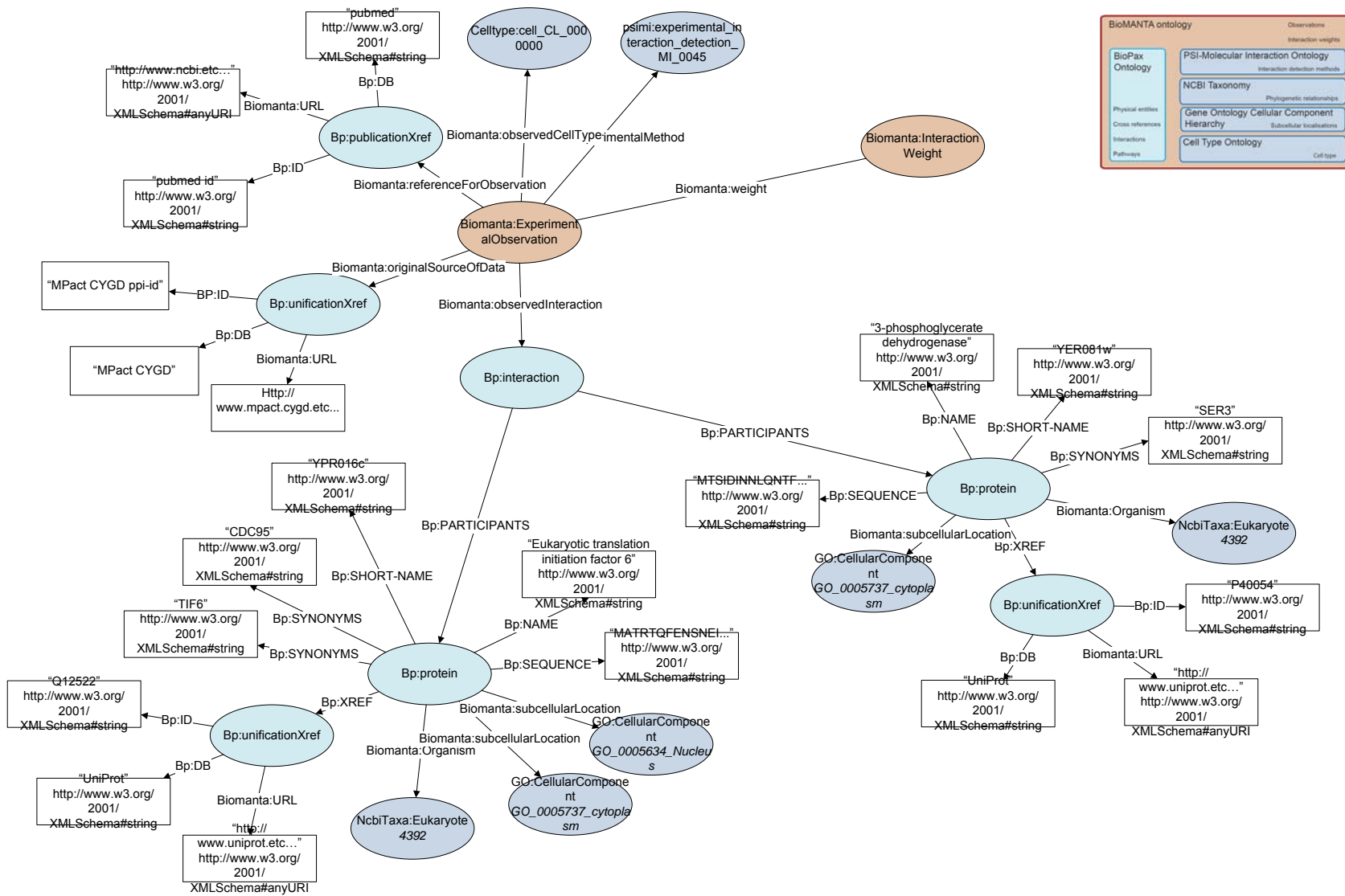
- Ontologies developed for specific uses
 - Describing pathways
 - Exchanging information
 - Controlling vocabulary for annotation
- Systems Biology necessitates overlapping domains and concepts
- Knowledge acquisition is costly and critical
 - reuse of existing ontology preferable

BioMANTA ontology

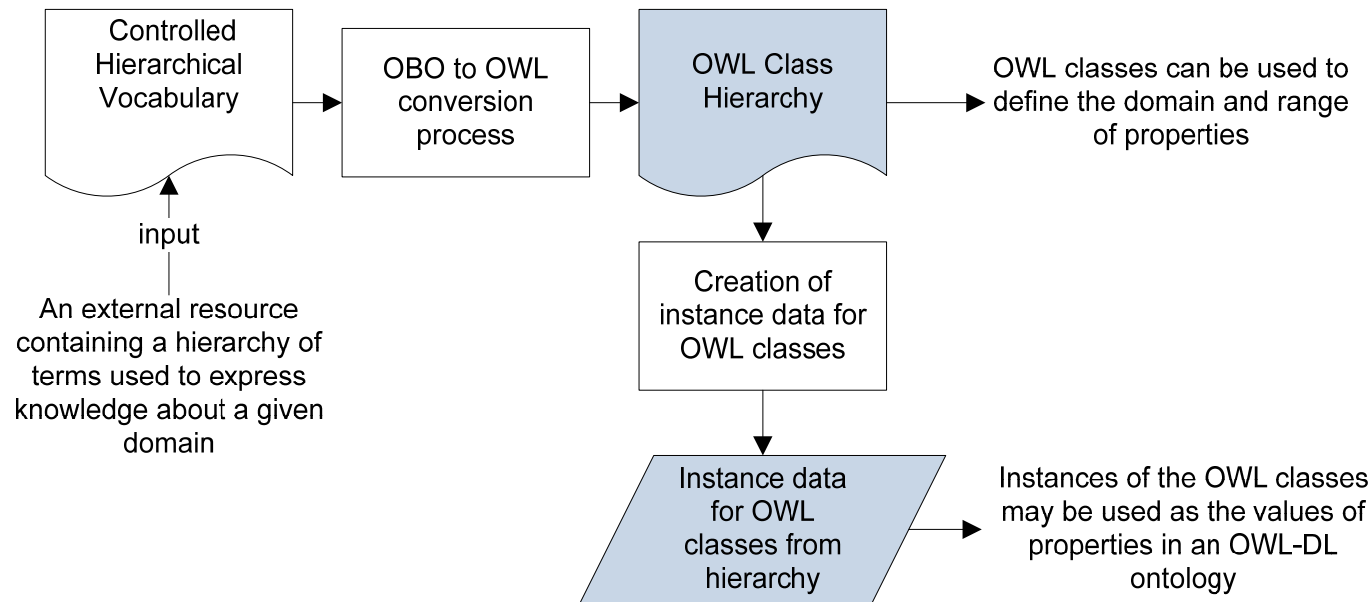


- OWL based ontology with imported modules from relevant ontologies
- Limit creation of new classes in BioMANTA ontology and use classes from existing ontologies where ever possible

BioMANTA Semantic Interactome Model



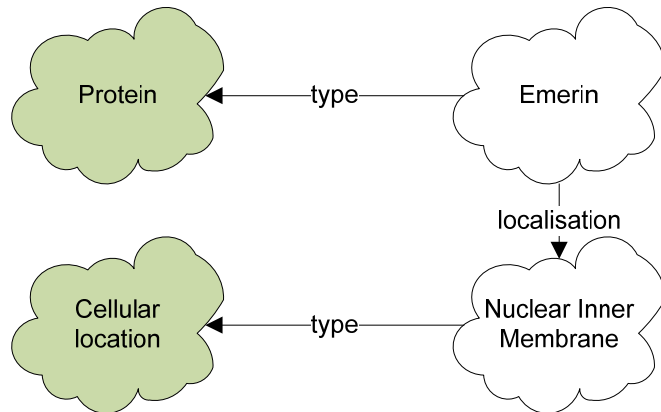
Ontology integration strategy



- Conversion and instancing of hierarchical controlled vocabulary
- Benefit: retention of meaning when framing expressions or instances of data

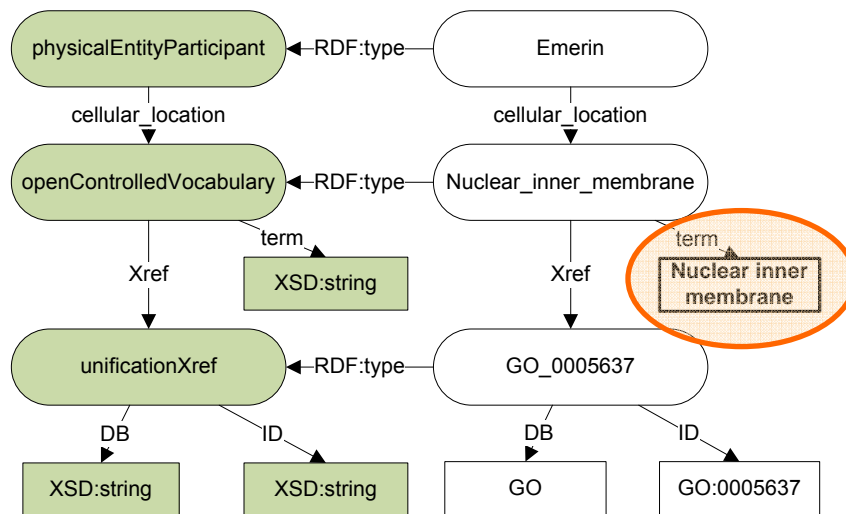
Example: Subcellular localisation

“The protein Emerin is localised to the nuclear inner membrane”



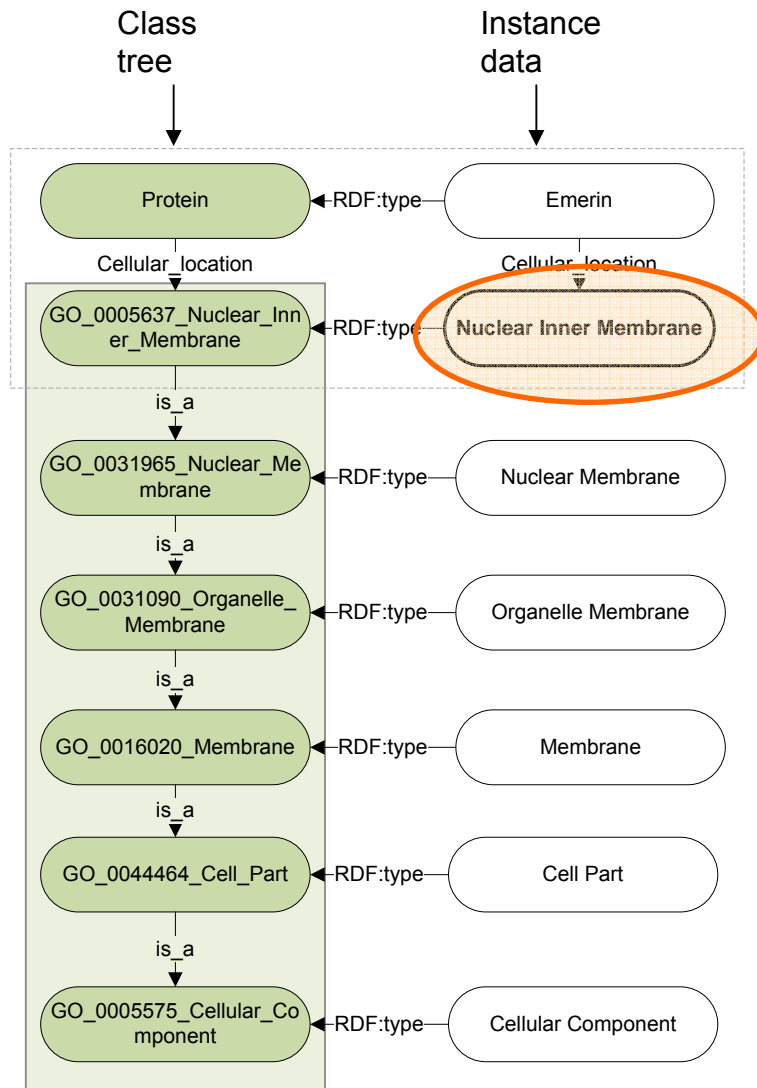
- Natural language expression

“The protein Emerin is localised to the nuclear inner membrane”



- BioPAX instance data expression
 - Text string
 - Dead end

Integrated Ontology in BioMANTA



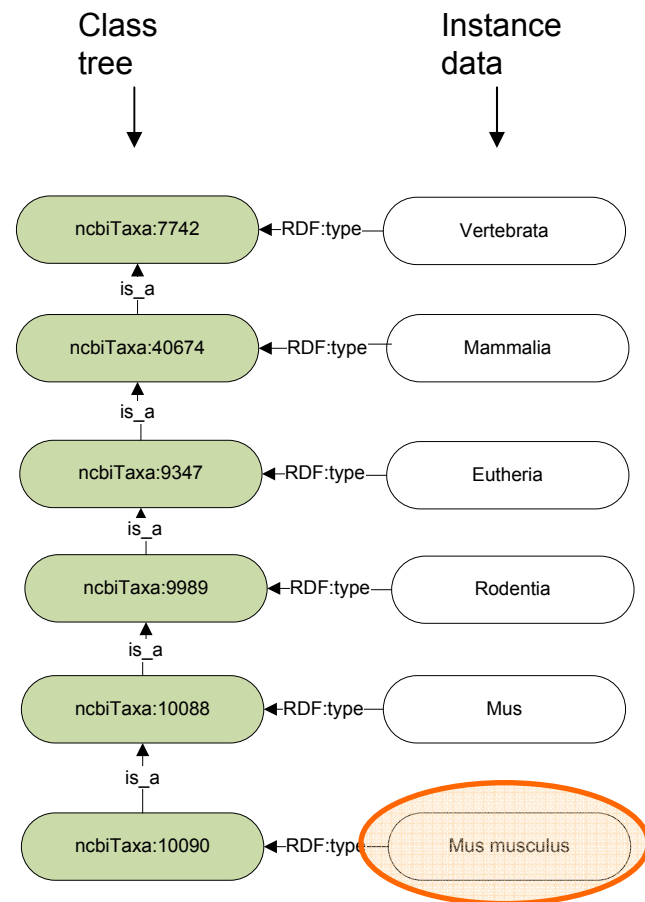
Statement that Emerin is localised to the nuclear inner membrane

Nuclear inner membrane is not merely a label, or annotation, it is embedded in a hierarchy of terms which capture additional meaning

Nuclear Inner Membrane *is_a* nuclear membrane and *is_a* membrane

Example: species and phylogeny

- How to reason about species relationships?
- Currently, labelling with text strings strips meaning from terms:
 - “mouse”; “mus musculus”; “10090”
 - meaning is lost – the string could be anything...
 - Nothing to say that “mouse” is mammal, which is basic background interpretation to a biologist
 - Class hierarchy (sub-class / super-class relationships) preserves meaning and captures assumptions



Conclusions

- Much biological knowledge is stored in controlled vocabularies inaccessible to machine reasoners
- Integration process salvages important collections of expert knowledge and brings them into SW format
- Creation of semantically rich labels enables inference
- Machine reasoners gain explicit access to the meaning of labeling terms – an improvement over text string inclusion used in other methods

Acknowledgements



Chris Bouton
Victor Farutin
Mike Schaffer
Fred Jerva

Pfizer Research and Technology
Center, Cambridge,
Massachusetts, US



IMB *Institute for Molecular Bioscience*

Jane Hunter
Andrew Newman
Imran Khan
Yuan-Fang Li

School of
ITEE, UQ

Mark Ragan
Kevin Burrage
Shoaib Sehgal

IMB & ARC Centre
of Excellence in
Bioinformatics

All members of Group Ragan @IMB