

Adapting resources to the Semantic Web: Experience with Entrez Gene

Satya S. Sahoo, Olivier Bodenreider, Kelly Zeng, Amit P. Sheth

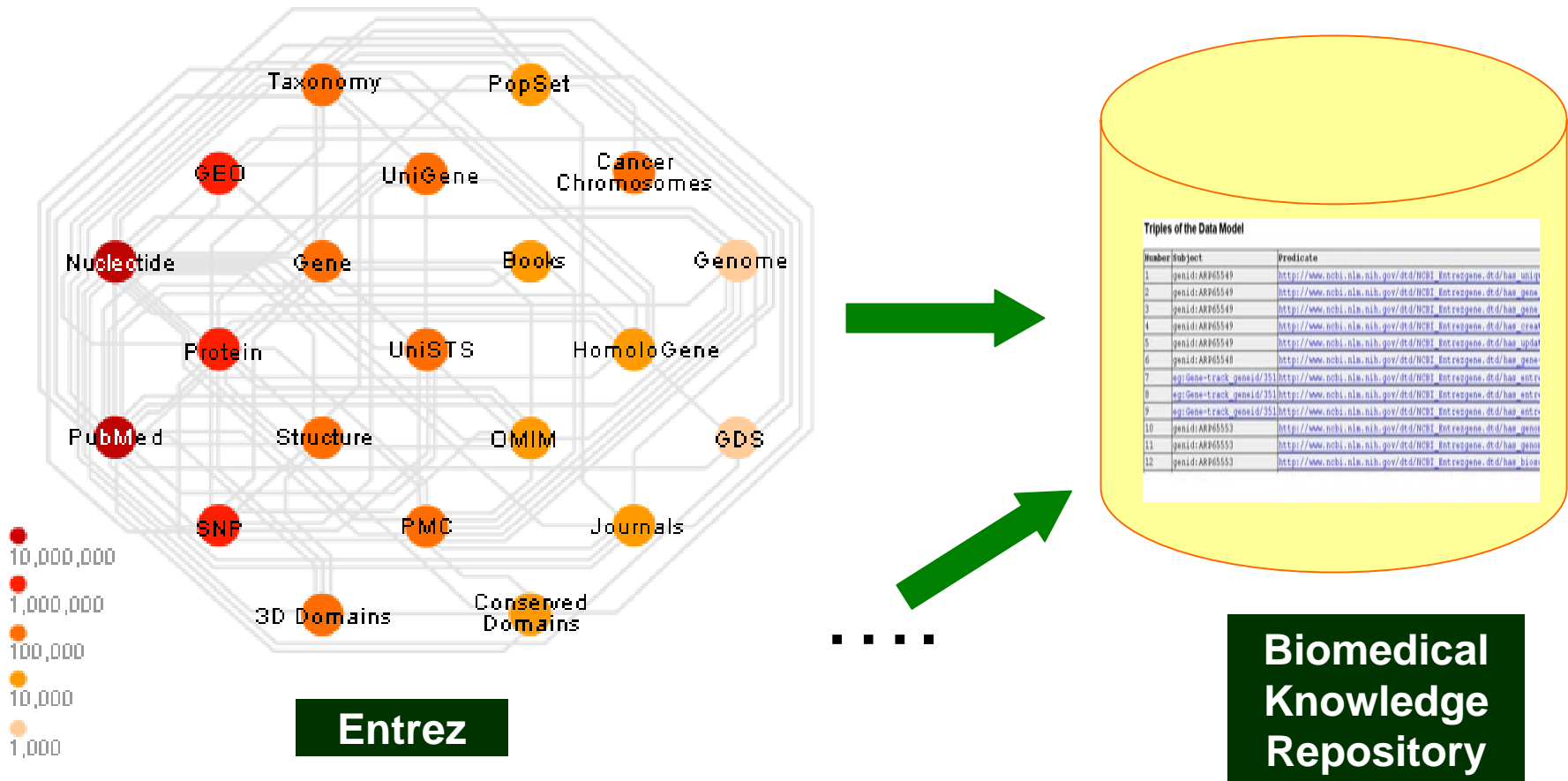
Presented By: Satya S Sahoo
<http://lsdis.cs.uga.edu/~satya/satya.html>

Presented at
W3C Semantic Web Health Care & Life Sciences Workshop
ISWC 2006, Athens, GA USA

Outline

- Motivation
- Materials and Methods
- Results
- Issues and challenges

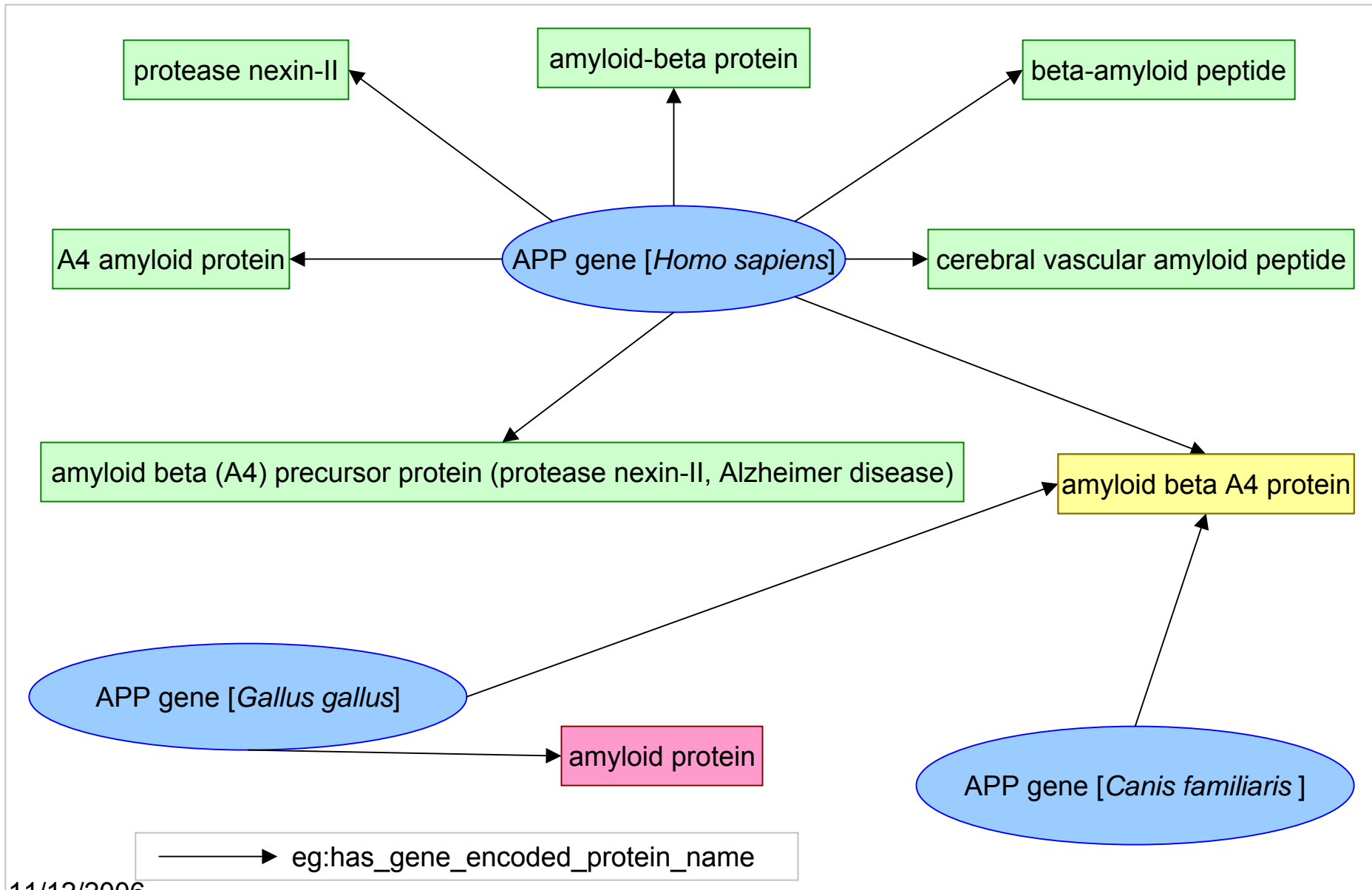
From supporting navigation to supporting knowledge processing



Relationships as first-class citizens

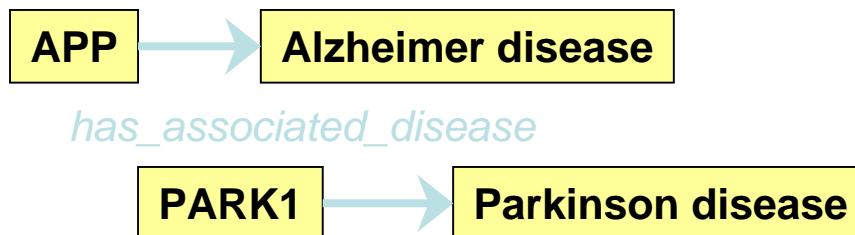
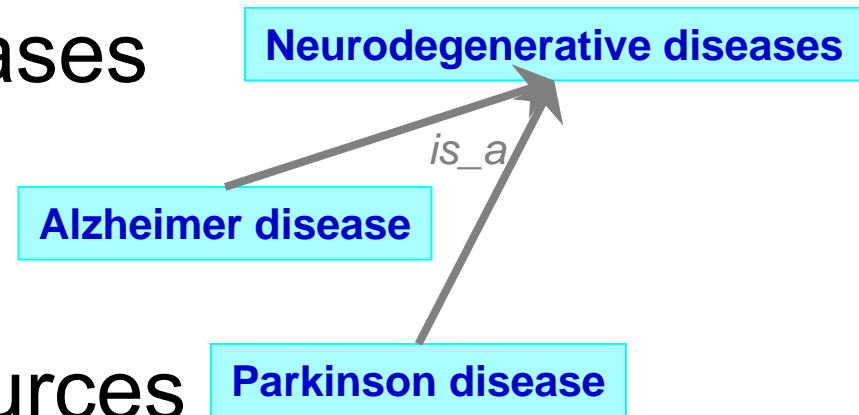
- Concentrate on the logical structure of data → **named relationships**
- Relationship-based complex query within a data resource
 - K-hop path query with specified end-points
 - Ranking of path query results using additional knowledge
- A formal model of relationships – currently a taxonomy

Integration *within* a resource



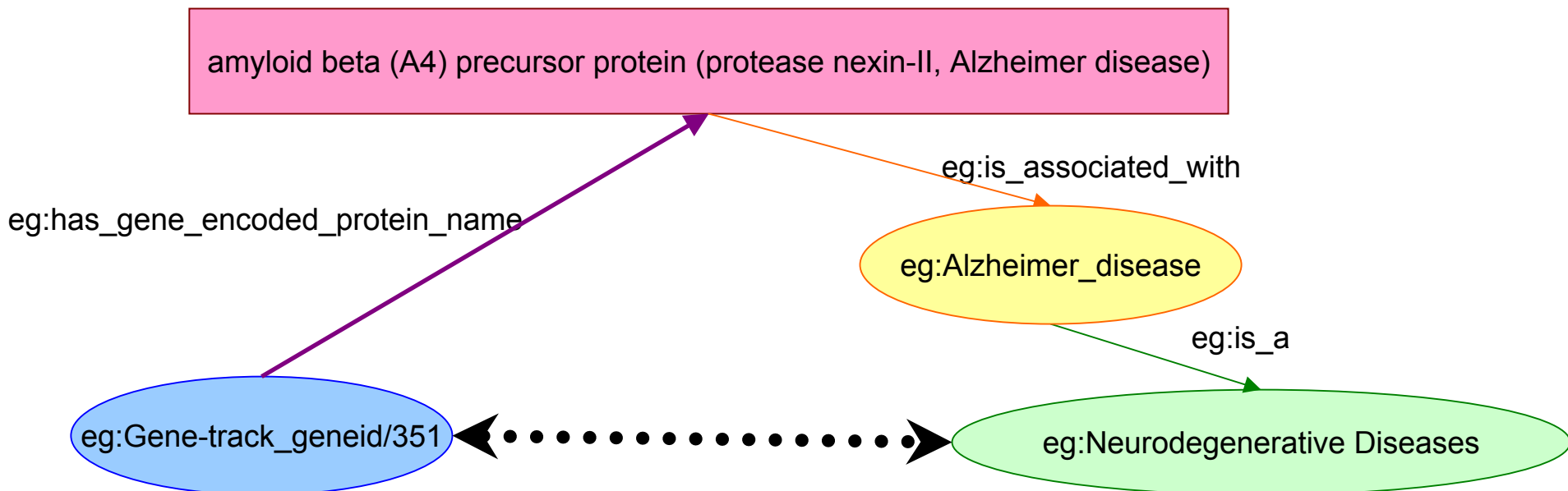
Integration *across* resources

- Transform additional resources into RDF
 - UMLS Metathesaurus
 - Other NCBI databases
 - Drug knowledge bases
 - ...
- Integrate resources
 - Query across resources



Inference

- Rules are objects that allow inference from RDF data [Alexander, N. et. al., Oracle]
- Oracle 10g allows the creation of rulebase based on RDFS (RDF Schema)



Outline

- Motivation
- Materials and Methods
- Results
- Issues and challenges

Entrez Gene Structure

NCBI Entrez Gene

My NCBI [Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books OMIM

Search Gene for APP amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease) Go Clear

Limits Preview/Index History Clipboard Details

Display Full Report Show 5 Send to

All: 1 Current Only: 1 Genes Genomes: 1 SNP GeneView: 1

1: **APP amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease)** [*Homo sapiens*] updated 26-Jul-2006
GeneID: 351 Primary source: [HGNC:620](#)

Summary

Official Symbol: APP **and Name:** amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease) **provided by** [HUGO Gene Nomenclature Committee](#)

See related: [HPRD:00100](#), [MIM:104760](#)

Gene type: protein coding

Gene name: APP

Gene description: amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease)

Entrez Gene Home

- Table Of Contents
- Summary
- Genomic regions, transcripts...
- Genomic context
- Bibliography
- HIV-1 protein interactions
- Interactions
- General gene information
- General protein information
- Reference Sequences
- Related Sequences
- Additional Links
- Links

General protein information

Names: amyloid beta A4 protein
protease nexin-II; A4 amyloid protein; amyloid-beta protein; beta-amyloid peptide; cerebral vascular amyloid peptide; amyloid beta (A4) precursor protein (protease nexin-II, Alzheimer disease)

Entrez Gene Structure

NCBI Entrez Gene

Search Gene for APP amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease) Go Clear

Display Full Report Show 5 Send to

All: 1 Current Only: 1 Genes Genomes: 1 SNP GeneView: 1

1: APP amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease) [Homo sapiens] updated 26-Jul-2006

GeneID: 351 Primary source: HGNC:620

Summary

Official Symbol: APP **and Name:** amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease) provided by HUGO

Gene Nomenclature Committee

See related: HPRD:00100, MIM:104760

Gene type: protein coding

Gene name: APP

Gene description: amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease)

has_protein_name

amyloid beta A4 protein

General protein information

Names: amyloid beta A4 protein
protease nexin-II; A4 amyloid protein; amyloid-beta protein; beta-amyloid peptide; cerebral vascular amyloid peptide; amyloid beta (A4) precursor protein (protease nexin-II, Alzheimer disease)

Entrez Gene Home
Table Of Contents
Summary
Genomic regions, transcripts...
Genomic context
Bibliography
HIV-1 protein interactions
Interactions
General gene information
General protein information
Reference Sequences
Related Sequences
Additional Links
Links

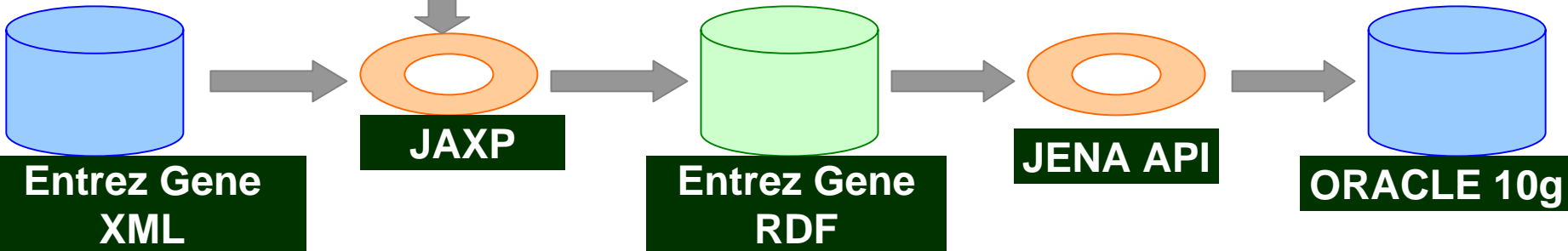
RDF Conversion Approach

- Preserve or enhance the information in native Entrez Gene
- Element tags to named relationships
 - 133 **element tags** (*attributes* considered as separate tags)
 - 105 **named relationships**
- Incorporate the notion of RDF containers to group logically similar values
- RDF model enables complex queries not possible on current repository

Implementation

```
<xsl:when test='$currNode="Entrezgene_track-info"'>  
  <xsl:element name="{ $ns }:has_entrezgene_track_info">  
    <xsl:if test="..../.* and ../.* and not (@*)">  
      <xsl:attribute name="rdf:parseType">Resource</xsl:attribute>  
    </xsl:if>  
  </xsl:element>  
</xsl:when>
```

XSLT stylesheet

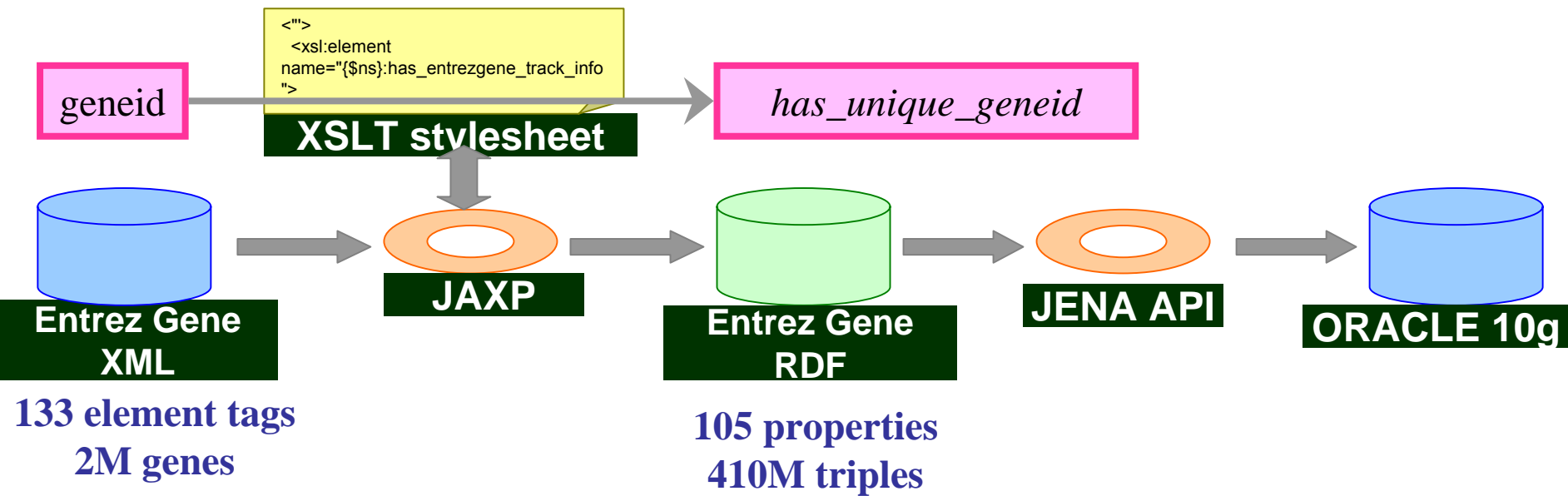


- Modular - Separates application code from transformation framework
- Extensible – Specific stylesheets may be used to for each of the Entrez databases
- Flexible – Changes in application logic or transformation logic are separate

Outline

- Motivation
- Materials and Methods
- Results
- Issues and challenges

Results



- 1 Record (geneid = 351) → 16,180 RDF triples
- Identified 3 candidate RDF containers → *Protein names, Gene reference synonyms, Organism reference synonyms*
- 50 GB Entrez Gene XML file → 39 GB RDF file
- Curation: Removed null literals and literals with leading space

Outline

- Motivation
- Materials and Methods
- Results
- Issues and challenges

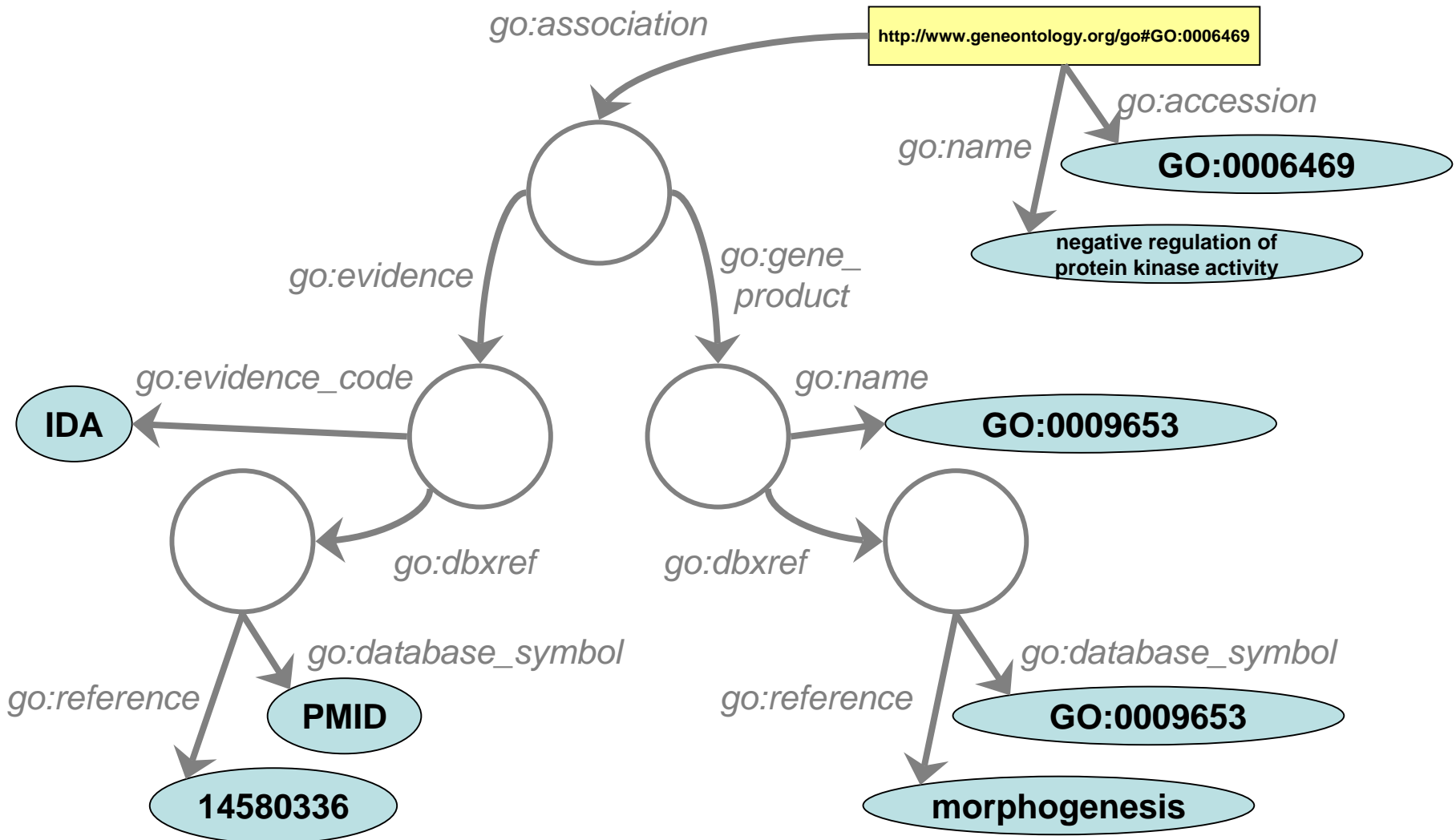
Separation of Data and Metadata - ongoing

- A gene record contains both:
 - **Metadata**: *date of creation, status, update date*
 - **Data**: *biological source organism, sequence intervals*
- Enable efficient query processing – two level filtering on resources
- **Difficulty**: context for metadata or data?

Issues and Challenges

- Blank nodes chain length?

RDF Blank nodes (Gene Ontology RDF)



Issues and Challenges

- Blank nodes chain length?
- Reconcile overlapping element tags in multiple data sources
- Nesting structure, bi-directionality of relations and, circularity need to be solved
- Relationships name should evolve with community participation

Issues and Challenges – Unique Identifier

- Identifier for biological entities is an issue of debate in the community
- Issues:
 - Can be dereferenced or not
 - Persistent or transient identifiers
- For now, we use the Entrez Gene DTD as the namespace
http://www.ncbi.nlm.nih.gov/dtd/NCBI_Entrezgene.dtd
- The possible candidates include:
 - LSID: Life Sciences Identifier
 - URI: NLM through UMLS and Entrez Gene

Further Information at:

http://esw.w3.org/topic/HCLSIG_BioRDF_Subgroup/Tasks/Entrez_Gene_to_RDF

Thank You

Bioinformatics Apps & Ontologies

- **GlycO**: A domain ontology for glycan structures, glycan functions and enzymes (embodying knowledge of the structure and metabolisms of glycans)
 - Contains 600+ classes and 100+ properties – describe structural features of glycans; unique population strategy
 - URL: <http://lsdis.cs.uga.edu/projects/glycomics/glyco>
- **ProPreO**: a comprehensive process Ontology modeling experimental proteomics
 - Contains 330 classes, 40,000+ instances
 - Models three phases of experimental proteomics* – Separation techniques, Mass Spectrometry and, Data analysis; URL: <http://lsdis.cs.uga.edu/projects/glycomics/propreo>
- Automatic semantic annotation of high throughput experimental data (**in progress**)
- Semantic Web Process with WSDL-S for semantic annotations of Web Services
 - <http://lsdis.cs.uga.edu/projects/glycomics/glyco/GlycOdoc2/>

More information at

- <http://lsdis.cs.uga.edu/projects/glycomics>