Converting biological information to the W3C Resource Description Framework (RDF): Experience with Entrez Gene

Presented By: Satya Sanket Sahoo

Mentor: Dr. Olivier Bodenreider

Outline

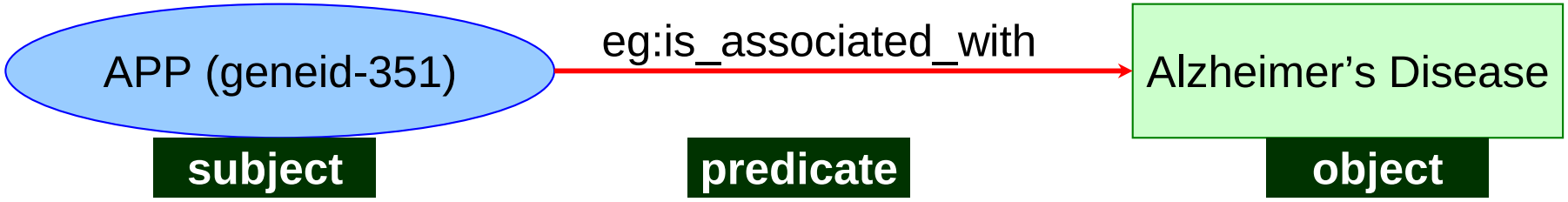
- Motivation
- RDF – Background
- Implementation technique
- Inference
- Unique identifiers
- Issues and challenges

Motivation: knowledge management

- Concentrate on the logical structure of data
- Explicit definition of terms and relationships
- Information integration – one universe for data from diverse background
- Inference: use existing knowledge to infer implicit knowledge

Resource Description Framework

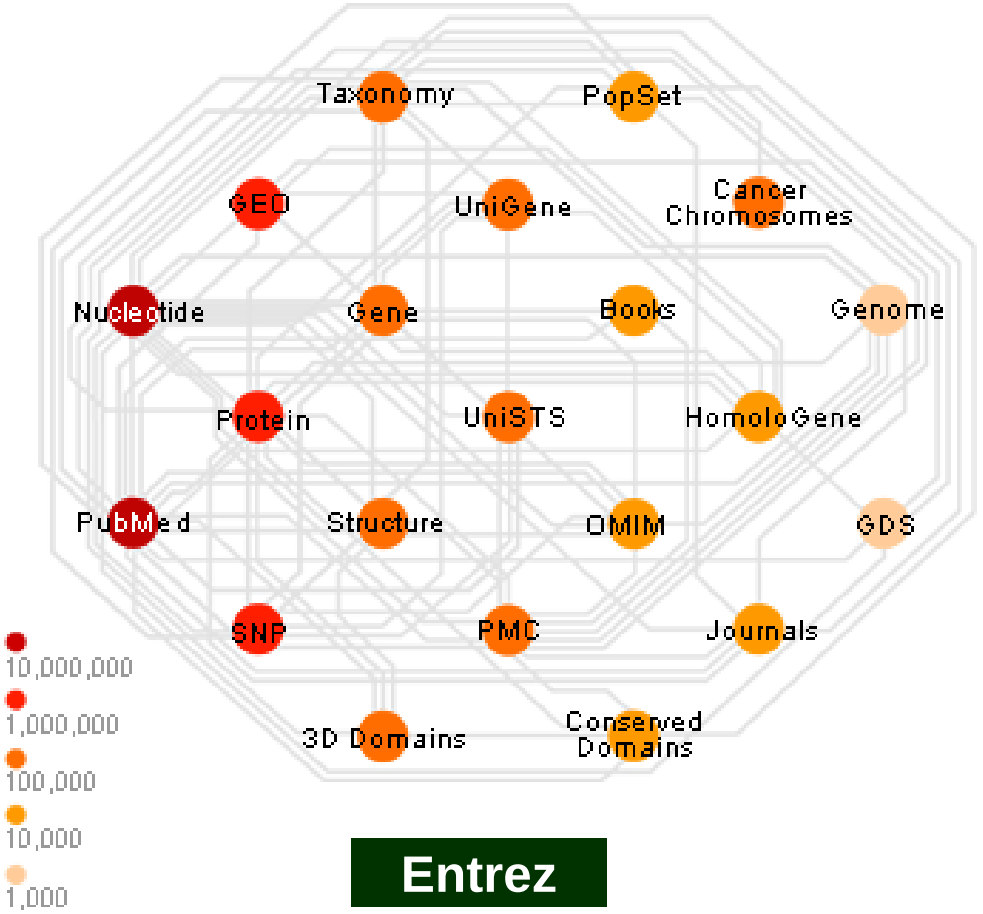
- All information represented as a ‘triple’



- Advantages include:
  - machine ‘understandable’
  - enables inference
  - represents the logical structure of the data
  - integration of data under one universe

Namespace - eg = http://www.ncbi.nlm.nih.gov/dtd/NCBI_Entrezgene.dtd/

RDF – contd.



Entrez

. . . .

Biomedical Knowledge Repository

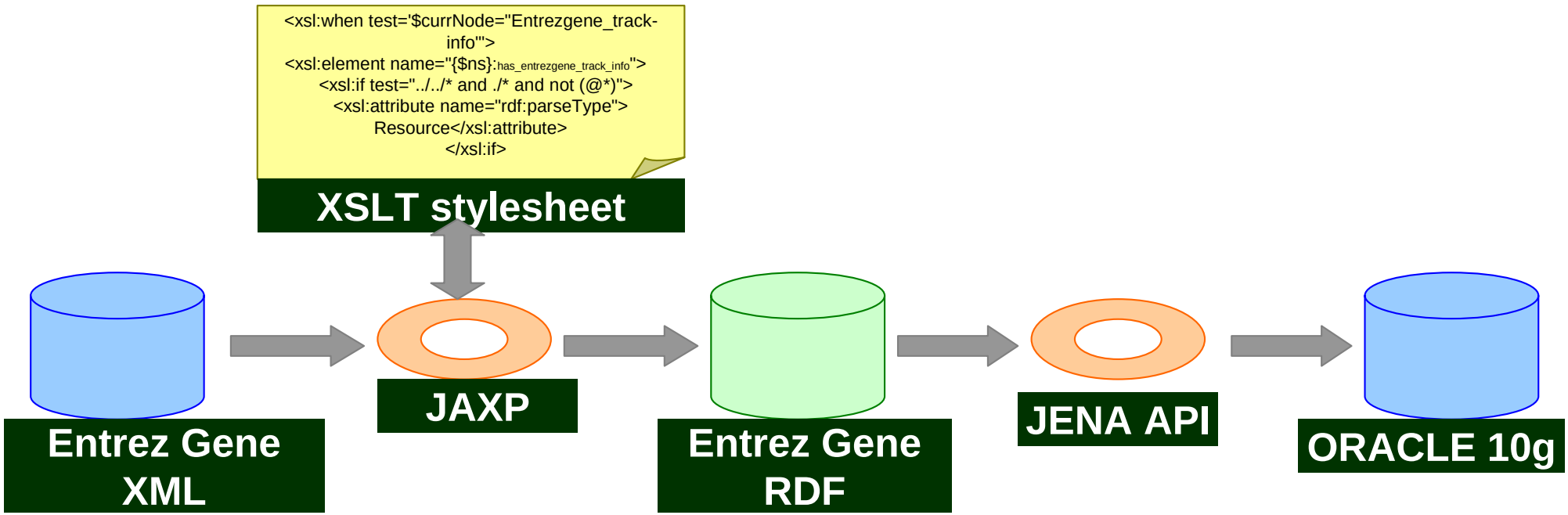
RDF – contd.

- RDF triples can be thought as normalized assertions
- Similar to normalization of text
- But, instead of **lexical resemblance** RDF triples enable **semantic resemblance**

Implementation: Entrez Gene XML to RDF

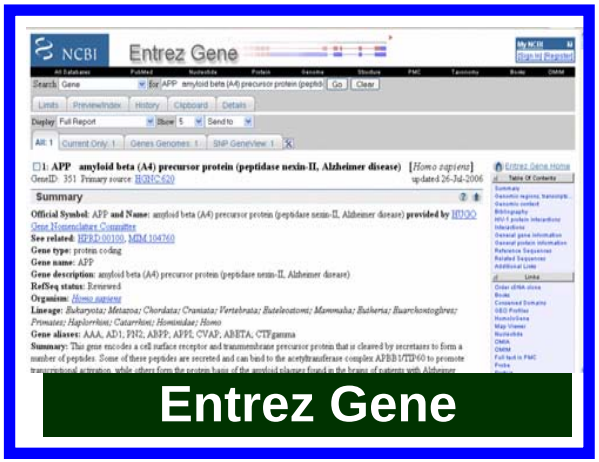
- Mapped element tags to more meaningful relations
- Started building an ontology of relationships
- Using XSLT stylesheet and XPath expressions converted XML to RDF
- The RDF reflects the nesting structure of terms in the Entrez gene records

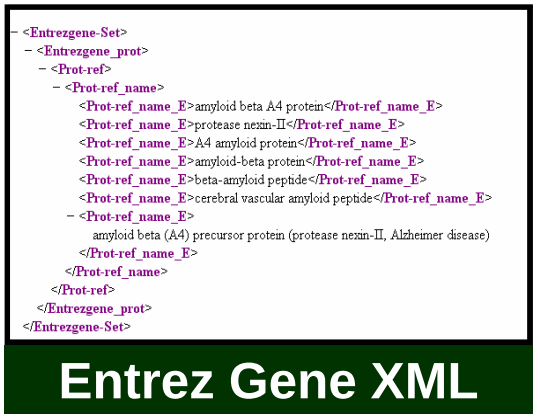
Implementation: Entrez Gene XML to RDF

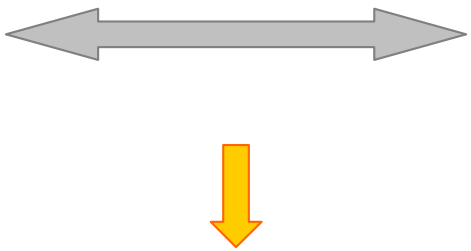


- Modular - Separates application code from transformation framework
- Extensible – specific stylesheets may be used to for each of the Entrez databases
- Flexible – changes in application logic or transformation logic are separate

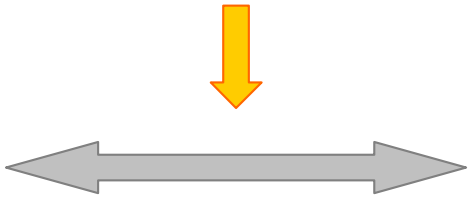
Implementation

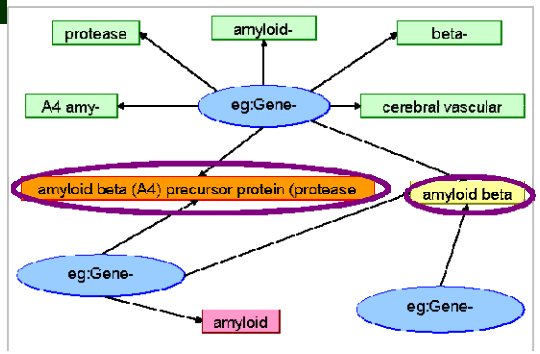




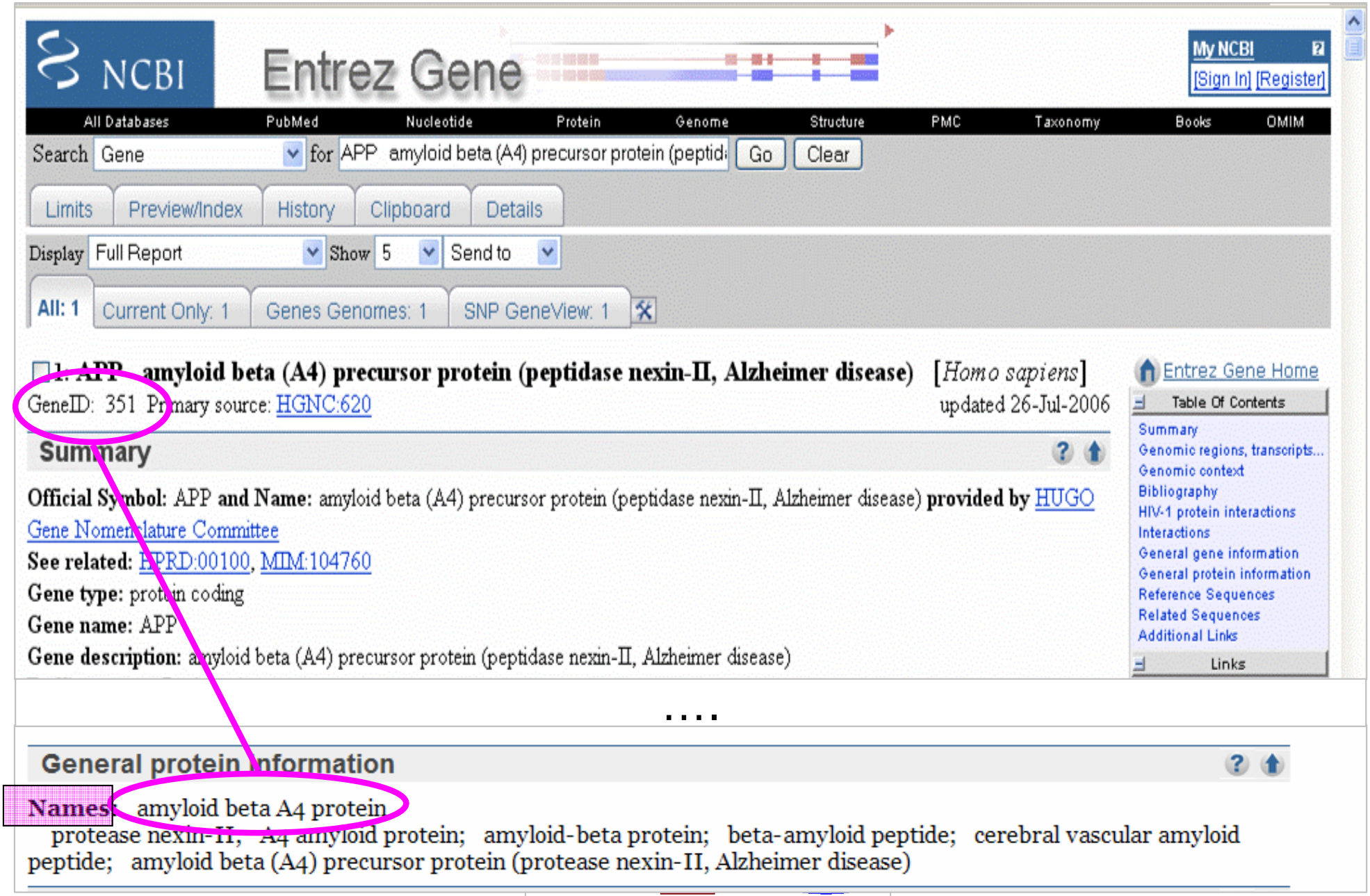


XSLT



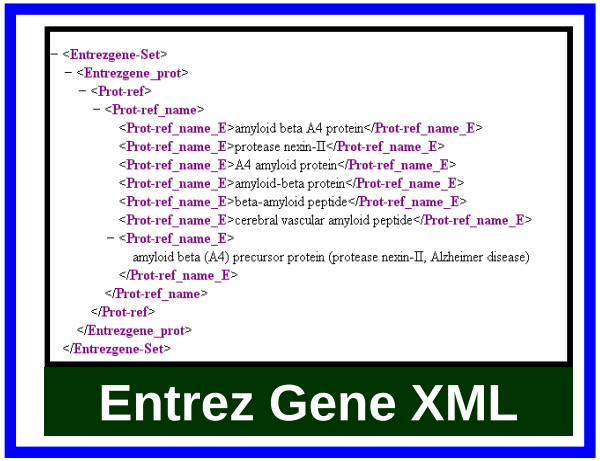


Web interface



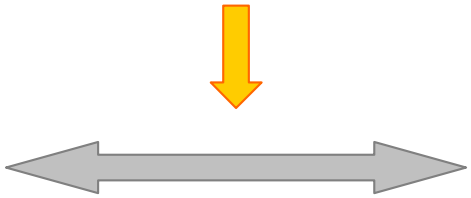
Implementation

Entrez Gene

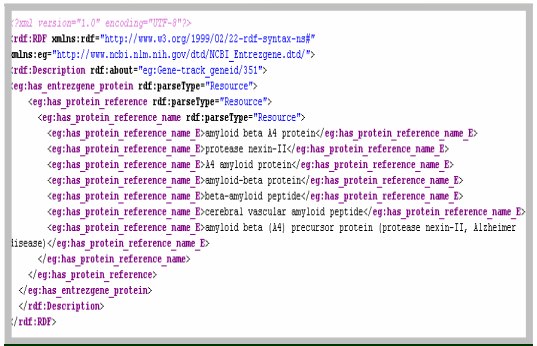


Entrez Gene XML

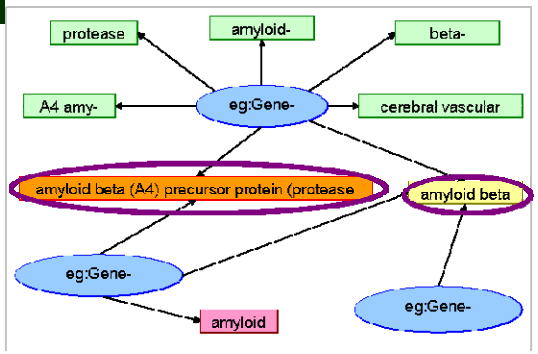
XSLT



Entrez Gene RDF graph



Entrez Gene RDF

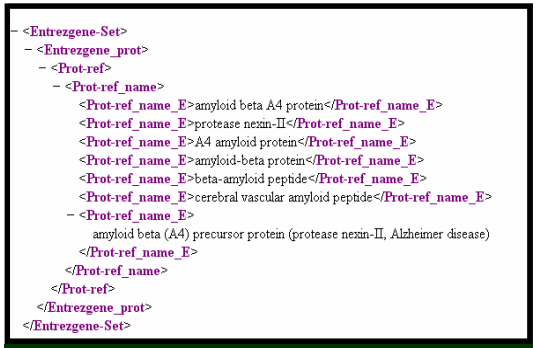


XML

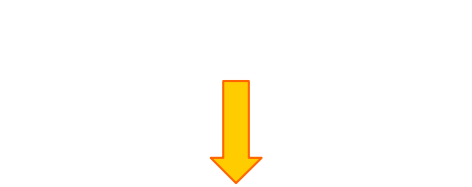
```
<?xml version="1.0"?>
<!DOCTYPE Entrezgene-Set PUBLIC "-//NLM//DTD NCBI-Entrezgene, 21st January 2005//EN" "NCBI_Entrezgene.dtd">
<Entrezgene-Set>
<Entrezgene>
<Entrezgene_track-info>
    <Gene-track>
      <Gene-track_geneid>351</Gene-track_geneid>
    </Gene-track>
</Entrezgene_track-info>
<Entrezgene_prot>
    <Prot-ref>
      <Prot-ref_name>
        <Prot-ref_name_E>amyloid beta A4 protein</Prot-ref_name_E>
        <Prot-ref_name_E>protease nexin-II</Prot-ref_name_E>
        <Prot-ref_name_E>A4 amyloid protein</Prot-ref_name_E>
        <Prot-ref_name_E>amyloid-beta protein</Prot-ref_name_E>
        <Prot-ref_name_E>beta-amyloid peptide</Prot-ref_name_E>
        <Prot-ref_name_E>cerebral vascular amyloid peptide</Prot-ref_name_E>
        <Prot-ref_name_E>amyloid beta (A4) precursor protein (protease nexin-II, Alzheimer disease)</Prot-ref_name_E>
      </Prot-ref_name>
    </Prot-ref>
  </Entrezgene_prot>
  </Entrezgene>
</Entrezgene-Set>
```

Implementation

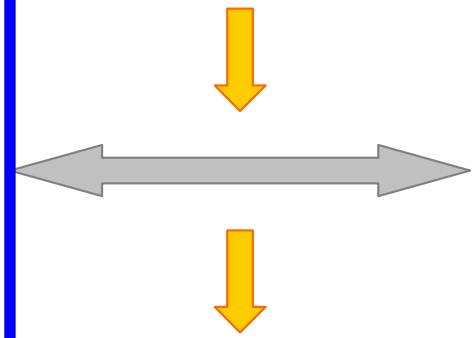
Entrez Gene

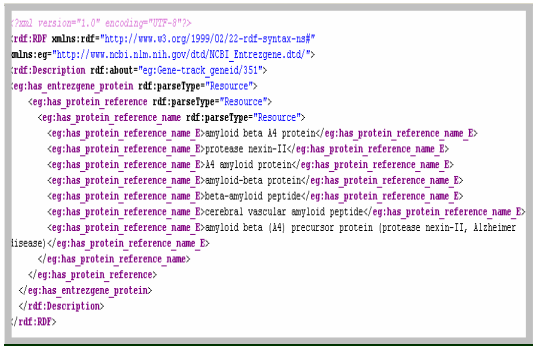


Entrez Gene XML

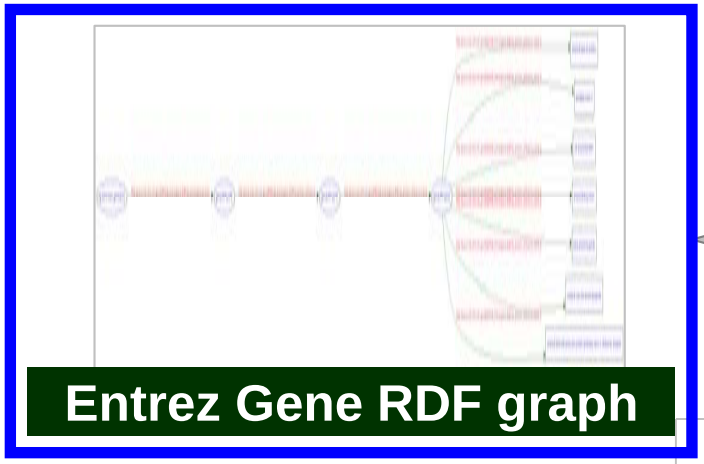


XSLT

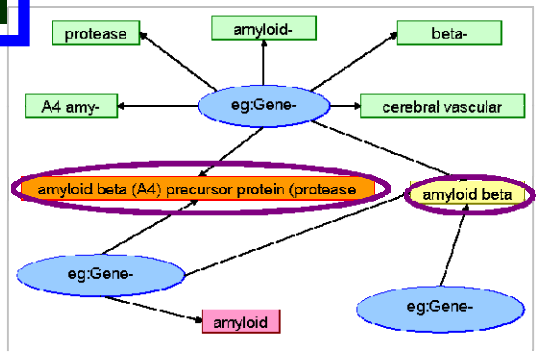




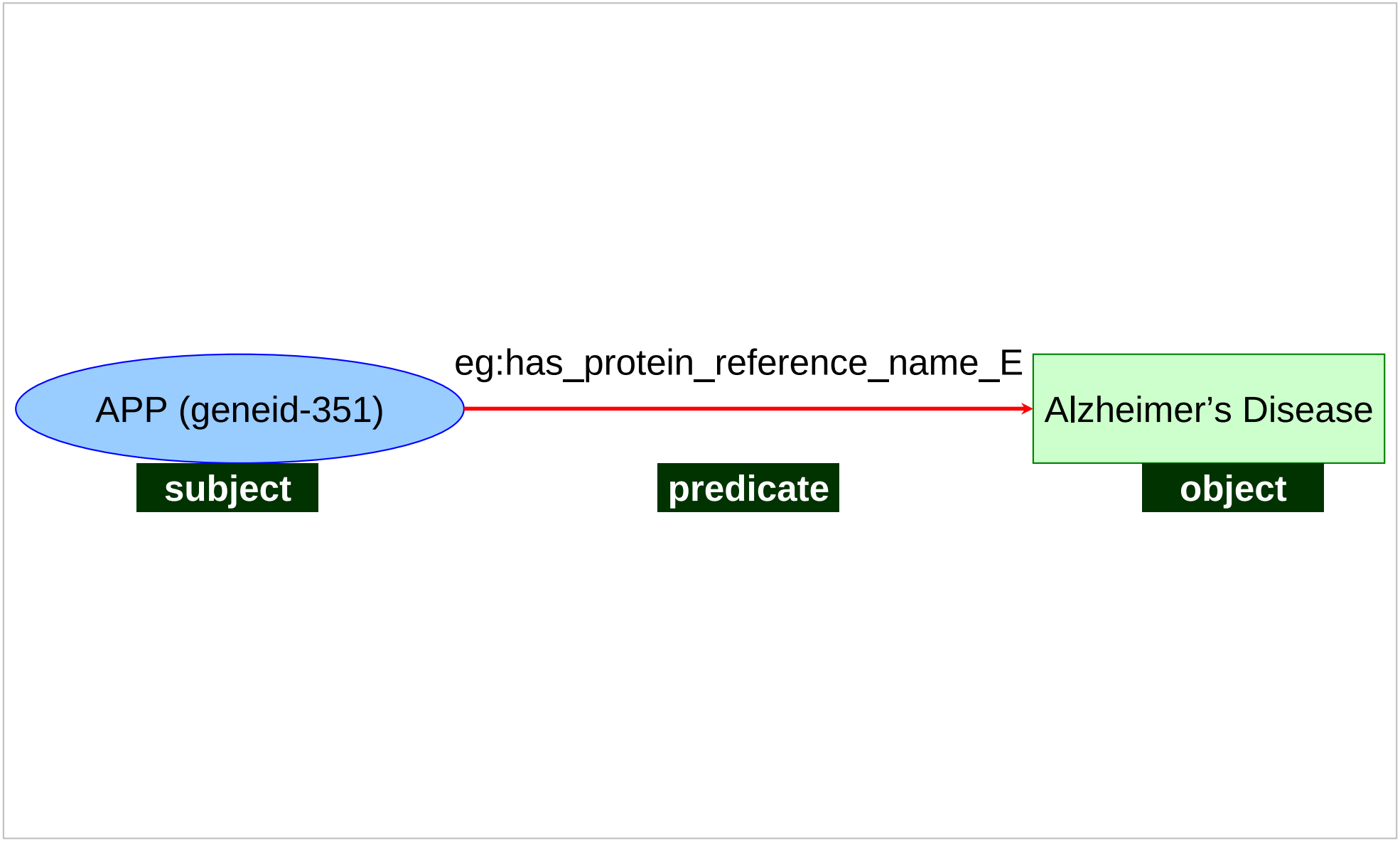
Entrez Gene RDF



Entrez Gene RDF graph



RDF Graph

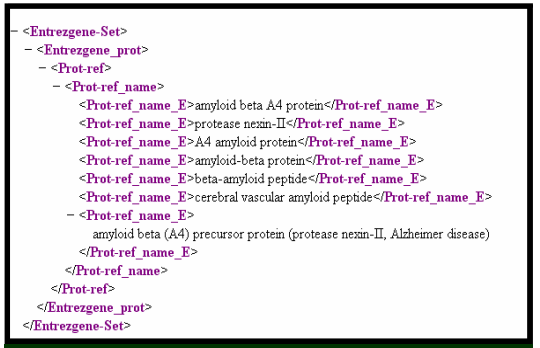

APP (geneid-351)
eg:has_protein_reference_name_E
Alzheimer's Disease
subject
predicate
object

RDF Graph

Entrez Gene RDF graph
(W3C Validator Site - http://www.w3.org/RDF/Validator/)

Implementation

Entrez Gene

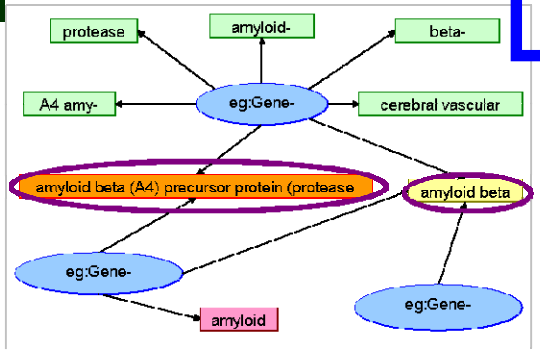


Entrez Gene XML

XSLT

Entrez Gene RDF graph

Entrez Gene RDF

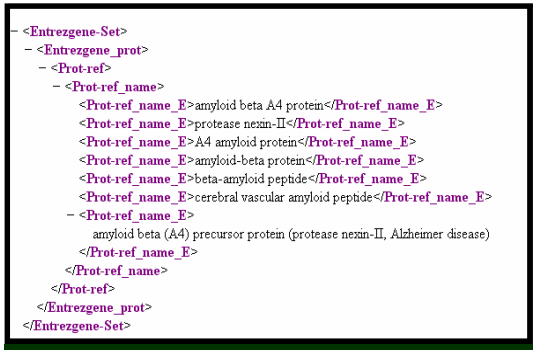


RDF

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:eg="http://www.ncbi.nlm.nih.gov/dtd/NCBI_Entrezgene.dtd/">
<rdf:Description rdf:about="eg:Gene-track_geneid/351">
<eg:has_entrezgene_protein rdf:parseType="Resource">
    <eg:has_protein_reference rdf:parseType="Resource">
      <eg:has_protein_reference_name rdf:parseType="Resource">
        <eg:has_protein_reference_name_E>amyloid beta A4 protein</eg:has_protein_reference_name_E>
        <eg:has_protein_reference_name_E>protease nexin-II</eg:has_protein_reference_name_E>
        <eg:has_protein_reference_name_E>A4 amyloid protein</eg:has_protein_reference_name_E>
        <eg:has_protein_reference_name_E>amyloid-beta protein</eg:has_protein_reference_name_E>
        <eg:has_protein_reference_name_E>beta-amyloid peptide</eg:has_protein_reference_name_E>
        <eg:has_protein_reference_name_E>cerebral vascular amyloid peptide</eg:has_protein_reference_name_E>
        <eg:has_protein_reference_name_E>amyloid beta (A4) precursor protein (protease nexin-II, Alzheimer
disease)</eg:has_protein_reference_name_E>
      </eg:has_protein_reference_name>
    </eg:has_protein_reference>
  </eg:has_entrezgene_protein>
  </rdf:Description>
</rdf:RDF>
```

Implementation

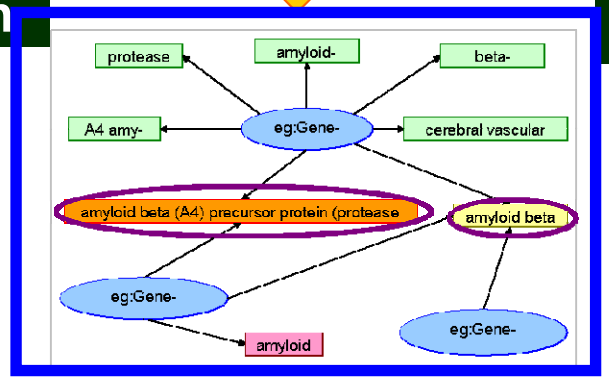
Entrez Gene



Entrez Gene XML

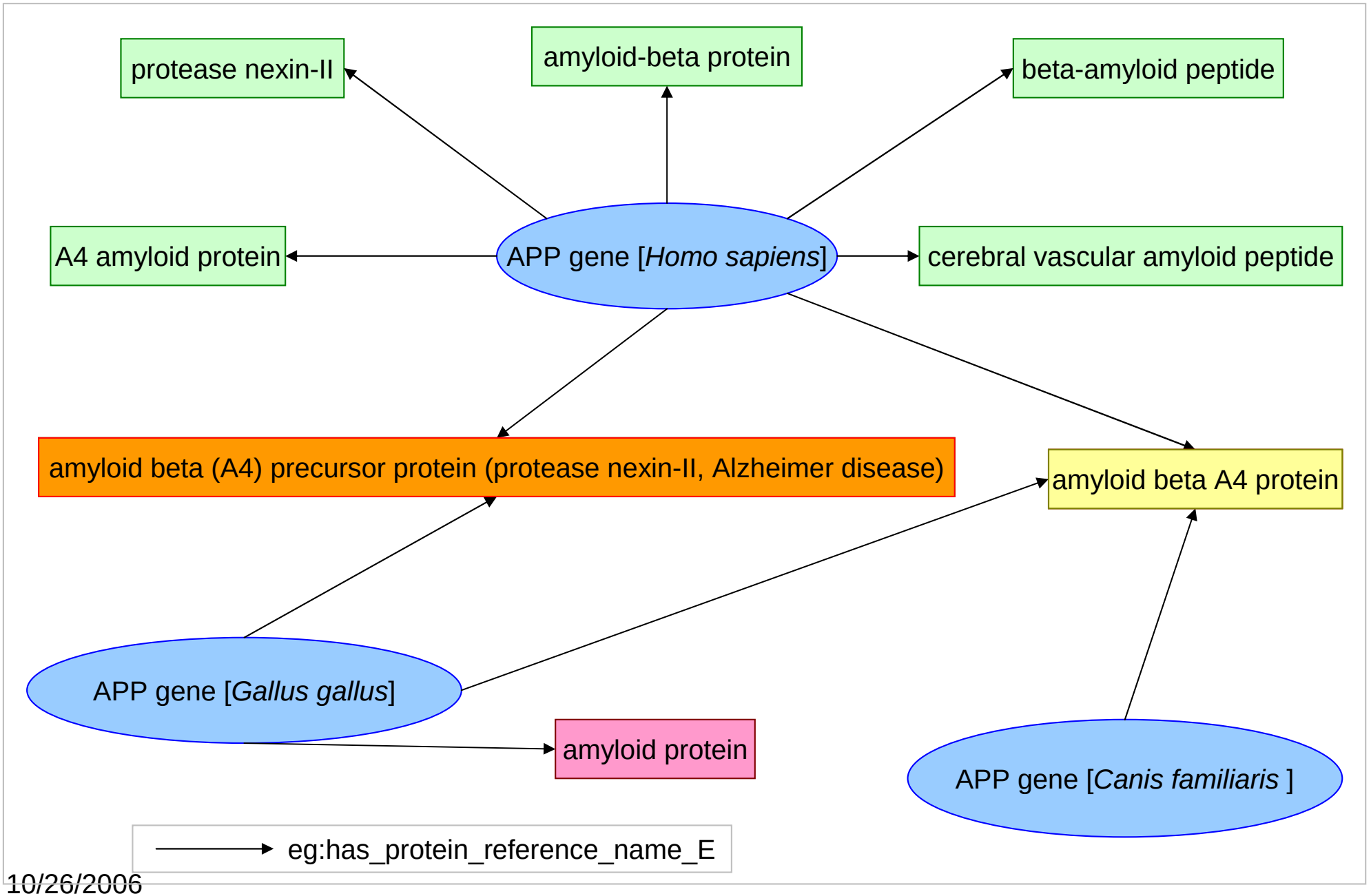
XSLT

Entrez Gene RDF graph



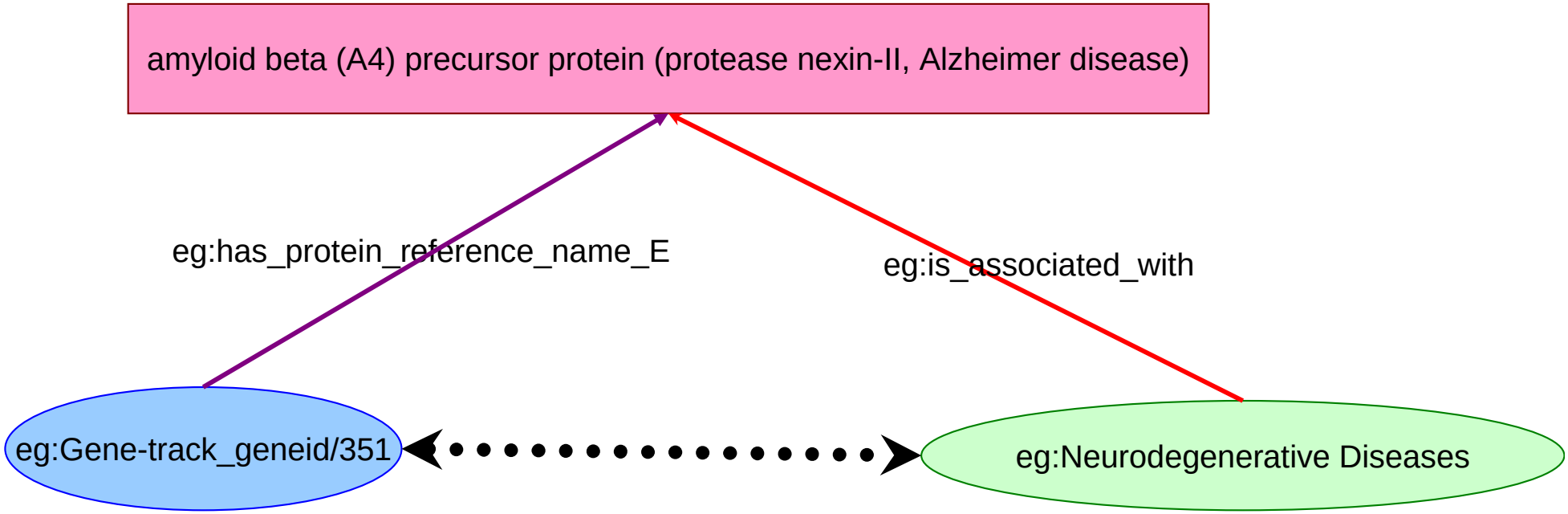
Entrez Gene RDF

Connecting different genes



Inference

- Rules are objects that allow inference from RDF data [1]
- Oracle 10g allows the creation of rulebase based on RDFS (RDF Schema)



Unique Identifier

- Identification of a resource uniquely
- Issues:
  - o Can be dereferenced or not
  - o Persistent or transient identifiers
- We use the Entrez Gene DTD as the namespace

  http://www.ncbi.nlm.nih.gov/dtd/NCBI_Entrezgene.dtd
- The possible candidates include:
  - o LSID: Life Sciences Identifier
  - o URI: NLM through UMLS and Entrez Gene

Issues and Challenges

- We implemented one of the multiple approaches for transformation
- Identifier for biological entities is an issue of debate in the community
- Nesting structure, bi-directionality of relations and, circularity need to be solved
- Evolve the form of relationships used as predicate in the triples

Special thanks to

- Kelly Zeng
- May Cheh
- Thomas C. Rindflesch
- Rob Logan
- Paul Lynch
- John Nyugen

References

1. Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T., “Entrez Gene: gene-centered information at NCBI”, *Nucleic Acids Res.* 2005 January 1; 33(Database Issue): D54–D58.
2. Resource Description Framework (RDF), http://www.w3.org/TR/2004/REC-rdf-primer-20040210/
3. Rindflesch, TC, Fiszman, M., “The Interaction of Domain Knowledge and Linguistic Structure in Natural Language Processing: Interpreting Hypernymic Propositions in Biomedical Text”, *Journal of Biomedical Informatics*. 2003;36(6):462-77.
4. XML Schema Language Transformation (XSLT), http://www.w3.org/TR/xslt
5. Alexander, N., Ravada S., “RDF Object Type and Reification in Oracle”—Technical White Paper (http://download-east.oracle.com/otndocs/tech/semantic_web/pdf/rdf_reification.pdf)
6. Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (http://www.ncbi.nlm.nih.gov/omim/)
7. BioRDF subgroup: http://esw.w3.org/topic/HCLSIG_BioRDF_Subgroup
8. McBride, B. 2002. Jena: A Semantic Web Toolkit. *IEEE Internet Computing* 6, 6 (Nov. 2002), 55-59.
9. XPath: http://www.w3.org/TR/xpath
10. Life Sciences Identifier (LSID) project: http://lsid.sourceforge.net/