Integrating Biomedical Knowledge
Experience with Entrez Gene
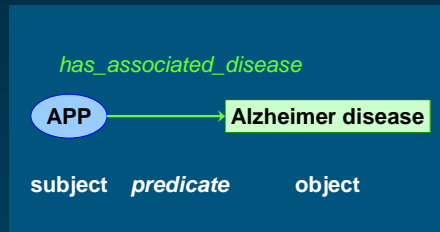
Olivier Bodenreider[1], Satya S. Sahoo[1,2] and Kelly Zeng[1]
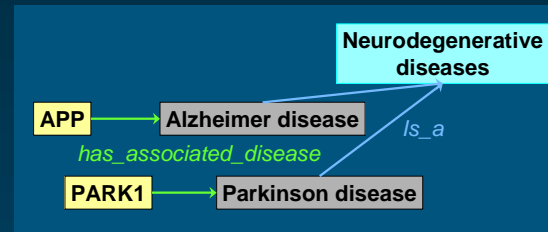
[1] LHNCBC, National Library of Medicine, Bethesda, MD
[2] LSDIS, University of Georgia, Athens, GA

Modern biomedical research is increasingly supported by information technologies. Biologists and physicians rely not only on the biomedical literature (e.g., *MEDLINE*), but also on the many knowledge bases available online. While these resources are undeniably valuable to humans, most of them are text-based and heterogeneous, and cannot be easily processed by computers.
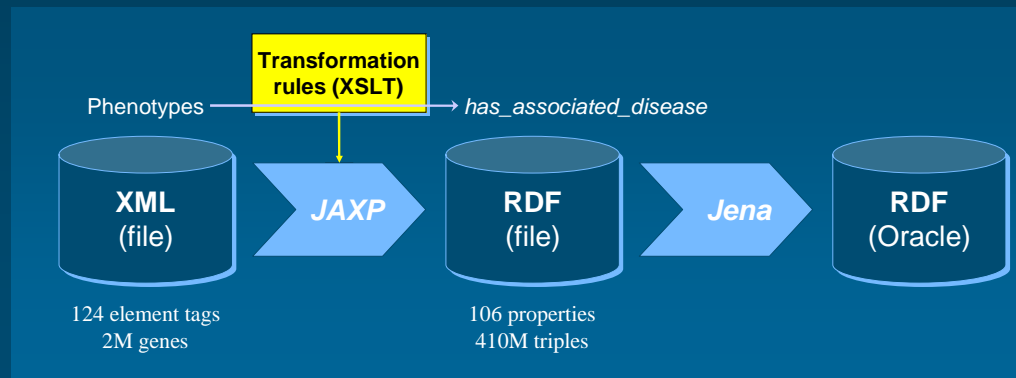
The *Biomedical Knowledge Repository* under development at the National Library of Medicine addresses these limitations. It can be understood as a specialized version of the Semantic Web. It consists of an extensive collection of assertions represented in a common format. Logical reasoners extend the capabilities of the repository by inferring new knowledge.

We converted NCBI's gene information resource *Entrez Gene* from its XML format to RDF, the *Resource Description Framework*. This transformation is not simply syntactic, but also semantic and was achieved through rules created manually and expressed in an XSLT *(Extensible Stylesheet Language Transformation)*.

◆

Once converted to RDF, Entrez Gene forms a graph and can be integrated seamlessly with other resources such as the *Medical Subject Headings* (MeSH). Hierarchies in MeSH can then be exploited to query genes, enabling researchers to formulate queries such as "Find all genes involved with neurodegenerative diseases".

Example of RDF triple

Example of RDF graph combining triples from Entrez Gene and MeSH

Converting Entrez Gene to RDF using an XSLT approach

124 element tags
2M genes

106 properties
410M triples

NIH Research Festival 2006

Contact: Olivier Bodenreider
olivier@nlm.nih.gov