

# Semantic Web for Life Science Data Representation and Integration

Kei Cheung, PhD

Yale Center for Medical Informatics



**CBB752, March 5, 2014, Yale University**



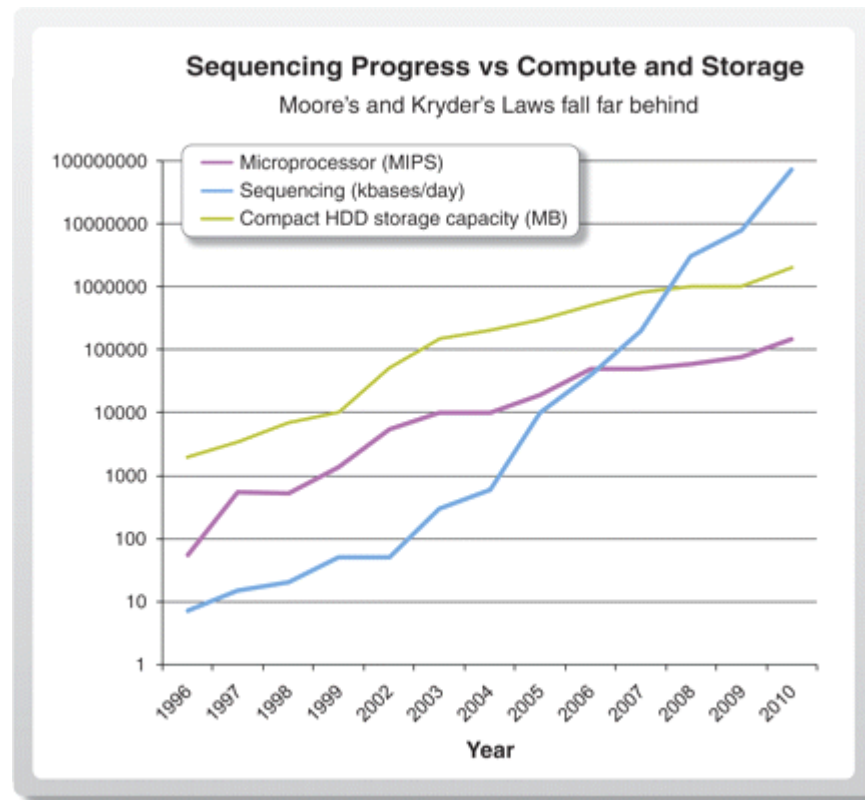
# in Science

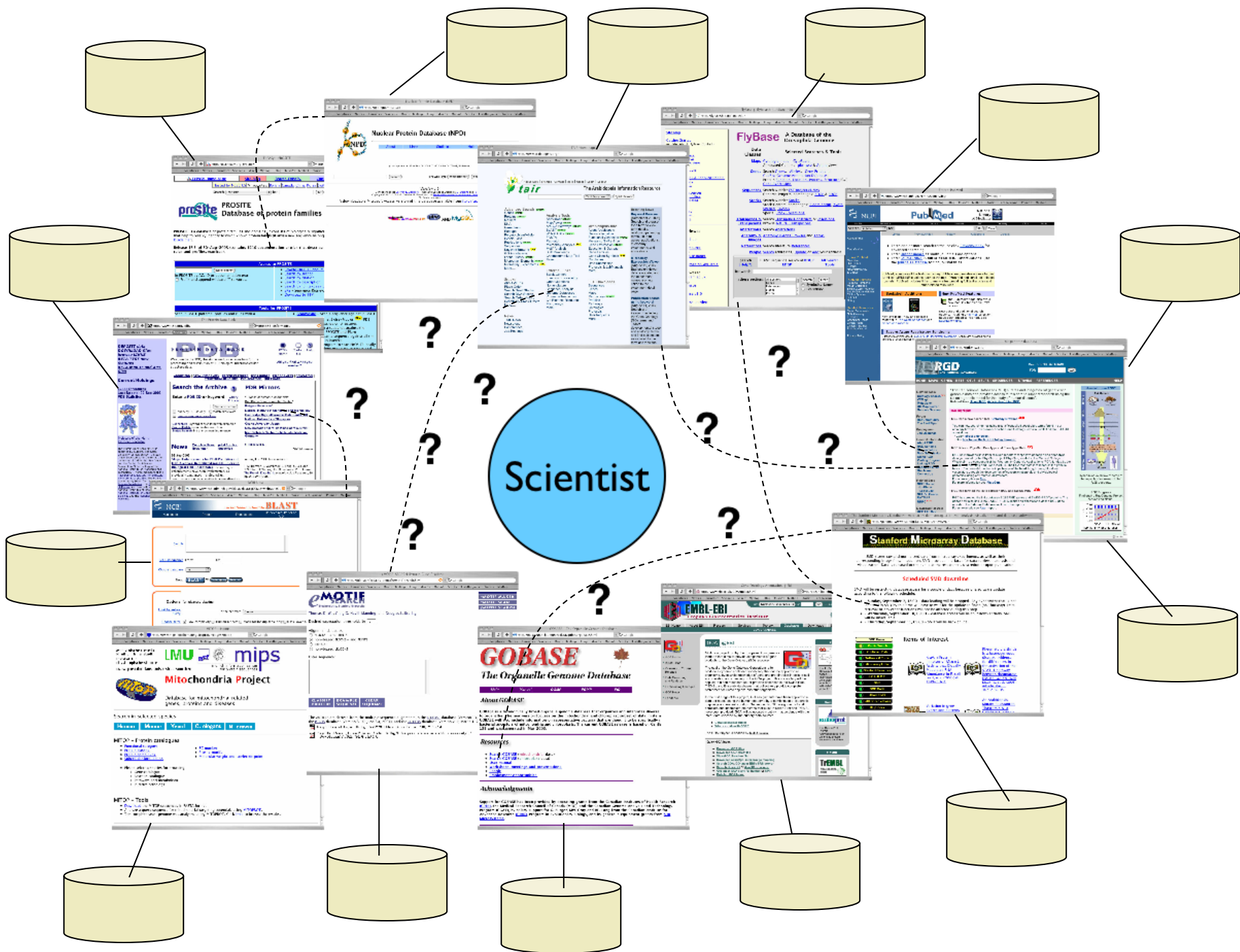


**(Nature Special Issue, September 2008)**



# Big Data in Genome Sciences





Can Google answer every question?!



# Problem with Google



Web [Show options...](#)

Results 1 - 10 of about 111,000 for kei cheung image

[Are you Kei Cheung?](#) Create your own profile on Google

[www.google.com/profiles](http://www.google.com/profiles) Help people find the right information when they search for you.

Image results for **kei cheung image** - [Report images](#)



[Kei Cheung - Email, Address, Phone numbers, everything! 123people.com](#)

123people finds photos related to **Kei Cheung** by using other search engines in real time.

The preview of the displayed **image** is associated with the original ...

[www.123people.com/s/kei+cheung](http://www.123people.com/s/kei+cheung) - [Share](#) [Print](#) [Close](#)

[Hoi Cheung - Email, Address, Phone numbers, everything! 123people.com](#)

123people never copies or stores any **image** files. If you are Hoi **Cheung** and ... Articles by **Kei-Hoi Cheung**. Fourth IEEE Symposium on Bioinformatics and ...

[www.123people.com/s/hoi+cheung](http://www.123people.com/s/hoi+cheung) - [Share](#) [Print](#) [Close](#)

[|kcheng|](#)

home.

[www.kchengimages.com/](http://www.kchengimages.com/) - [Cached](#) - [Similar](#) - [Share](#) [Print](#) [Close](#)

[VINDA - Kei Cheung Industries & Trading Limited](#)

It is a "everyday life paper professional" **image** in cinsumer minds with more careness and

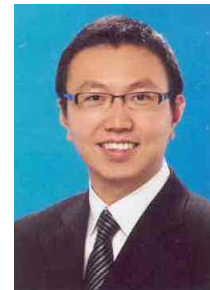
... Copyrights **Kei Cheung** Industries & Trading Limited 2008. ...

[www.keicheung.com.hk/VINDA.html](http://www.keicheung.com.hk/VINDA.html) - [Cached](#) - [Share](#) [Print](#) [Close](#)



I'm NOT a company!

其昌實業貿易有限公司  
Kei Cheung Industries & Trading Limited



Kei (Hui) Cheung  
Not me!



Kei (Hoi) Cheung  
(20 years ago)



Kei (Hoi) Cheung  
(more recent)

Find the most recent image  
of the person "Kei Hoi Cheung"

# Semantic Google

Google obama   Sign in

Web News **Images** Videos Shopping More Search tools Safe Search 

Size **Black and white** Photo **Jan 1, 1990 – Feb 27, 2014** Usage rights More tools Clear



The image grid displays 28 black and white photographs of Barack Obama, illustrating semantic search results. The images include:

- A man in a library reaching for a book.
- A formal portrait of Barack Obama.
- A profile shot of Barack Obama with his hand to his chin.
- A high-contrast, stylized portrait of Barack Obama.
- A film strip showing multiple frames of Barack Obama wearing a hat.
- A close-up portrait of a smiling Barack Obama.
- A portrait of a young Barack Obama.
- A portrait of a young Barack Obama.
- A group photo of a basketball team with "BARBARA" and "KUMAA" visible on their jerseys.
- A portrait of a young man with a mustache.
- A portrait of a young Barack Obama.
- A portrait of a young Barack Obama with the caption "Barry Obama".
- A photo of a woman with the caption "GO PLAY HOOP".
- A portrait of a young Barack Obama.
- A portrait of Barack Obama with the caption "Barry Obama".
- A portrait of Barack Obama.
- A portrait of Barack Obama wearing a hat.
- A portrait of a woman.
- A portrait of Barack Obama shouting.
- A portrait of a young Barack Obama.
- A portrait of Barack Obama in a suit.
- A portrait of Barack Obama in a suit.
- A portrait of a young Barack Obama.
- A portrait of Barack Obama with the text: "THIS IS OUR FIRST TASK, CARING FOR OUR CHILDREN. IT'S OUR FIRST JOB. IF WE DON'T GET THAT RIGHT, WE DON'T GET ANYTHING RIGHT. THAT'S HOW AS A SOCIETY, WE WILL BE JUDGED."
- A full-body photo of Barack Obama in a suit.



## Thing > CreativeWork > Dataset

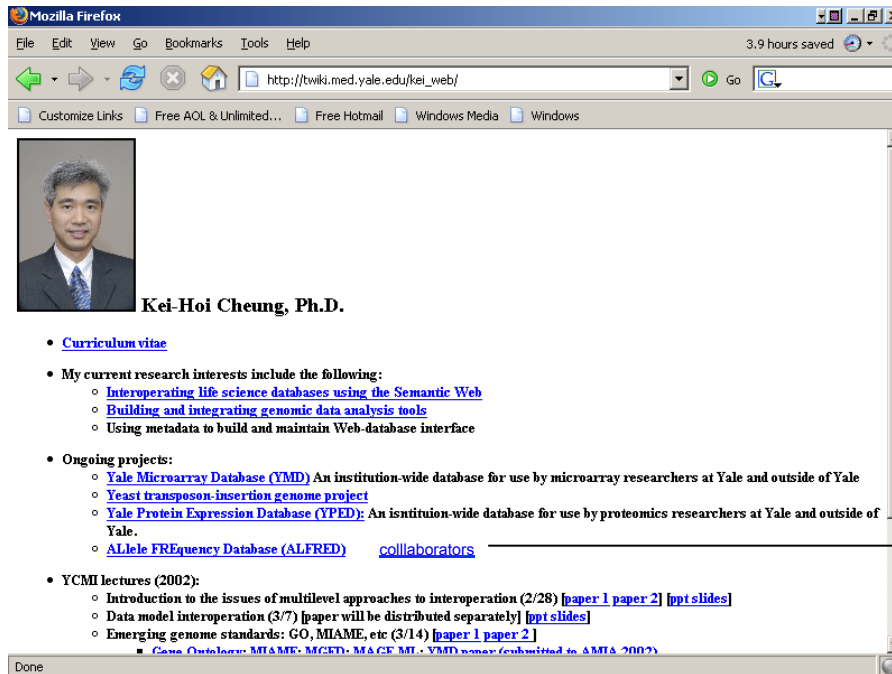
A body of structured information describing some topic(s) of interest.

Property	Expected Type	Description
<b>Properties from <u>Thing</u></b>		
<u>additionalType</u>	URL	An additional type for the item, typically used for adding more specific types from external vocabularies in microdata syntax. This is a relationship between something and a class that the thing is in. In RDFa syntax, it is better to use the native RDFa syntax – the 'typeof' attribute – for multiple types. Schema.org tools may have only weaker understanding of extra types, in particular those defined externally.
<u>alternateName</u>	Text	An alias for the item.
<u>description</u>	Text	A short description of the item.
<u>image</u>	URL	URL of an image of the item.
<u>name</u>	Text	The name of the item.
<u>sameAs</u>	URL	URL of a reference Web page that unambiguously indicates the item's identity. E.g. the URL of the item's Wikipedia page, Freebase page, or official website.
<u>url</u>	URL	URL of the item.
<b>Properties from <u>CreativeWork</u></b>		
<u>about</u>	<u>Thing</u>	The subject matter of the content.
<u>accessibilityAPI</u>	Text	Indicates that the resource is compatible with the referenced accessibility API ( <a href="#">WebSchemas wiki lists possible values</a> ).
<u>accessibilityControl</u>	Text	Identifies input methods that are sufficient to fully control the described resource ( <a href="#">WebSchemas wiki lists possible values</a> ).
<u>accessibilityFeature</u>	Text	Content features of the resource, such as accessible media, alternatives and supported enhancements for accessibility ( <a href="#">WebSchemas wiki lists possible values</a> ).
<u>accessibilityHazard</u>	Text	A characteristic of the described resource that is physiologically dangerous to some users. Related to WCAG 2.0 guideline 2.3. ( <a href="#">WebSchemas wiki lists possible values</a> )
<u>accountablePerson</u>	<u>Person</u>	Specifies the Person that is legally accountable for the CreativeWork.
<u>aggregateRating</u>	<u>AggregateRating</u>	The overall rating, based on a collection of reviews or ratings, of the item.
<u>alternativeHeadline</u>	Text	A secondary title of the CreativeWork.

# Problems of the Current Web

- Lack of use of global names/identifiers
- Lack of links and link semantics
- Lack of data standards and semantics


# Lack of Links and Link Semantics



Mozilla Firefox  
File Edit View Go Bookmarks Tools Help 3.9 hours saved

http://twiki.med.yale.edu/kei\_web/

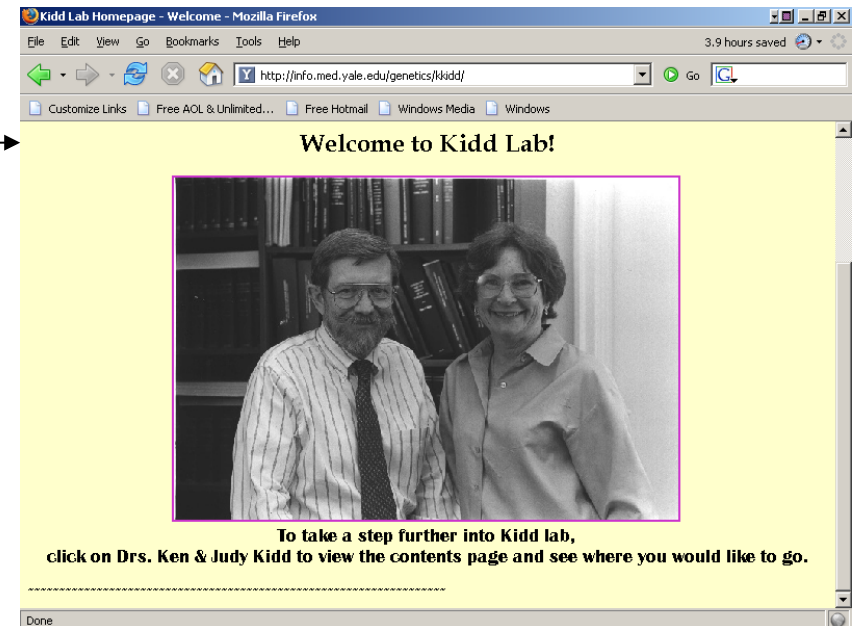
Customize Links Free AOL & Unlimited... Free Hotmail Windows Media Windows



**Kei-Hoi Cheung, Ph.D.**

- [Curriculum vitae](#)
- My current research interests include the following:
  - [Interoperating life science databases using the Semantic Web](#)
  - [Building and integrating genomic data analysis tools](#)
  - [Using metadata to build and maintain Web-database interface](#)
- Ongoing projects:
  - [Yale Microarray Database \(YMD\)](#) An institution-wide database for use by microarray researchers at Yale and outside of Yale
  - [Yeast transposon-insertion genome project](#)
  - [Yale Protein Expression Database \(YPED\)](#): An institution-wide database for use by proteomics researchers at Yale and outside of Yale.
  - [ALlele FREquency Database \(ALFRED\)](#) [collaborators](#)
- YCMI lectures (2002):
  - Introduction to the issues of multilevel approaches to interoperation (2/28) [[paper 1](#)] [[paper 2](#)] [[ppt slides](#)]
  - Data model interoperation (3/7) [paper will be distributed separately] [[ppt slides](#)]
  - Emerging genome standards: GO, MIAME, etc (3/14) [[paper 1](#)] [[paper 2](#)]
  - [Gene Ontology: MIAME, MGED, MAGE ML, YMD paper \(submitted to SMTA 2002\)](#)

Done




Kidd Lab Homepage - Welcome - Mozilla Firefox  
File Edit View Go Bookmarks Tools Help 3.9 hours saved

http://info.med.yale.edu/genetics/kidd/

Customize Links Free AOL & Unlimited... Free Hotmail Windows Media Windows

**Welcome to Kidd Lab!**

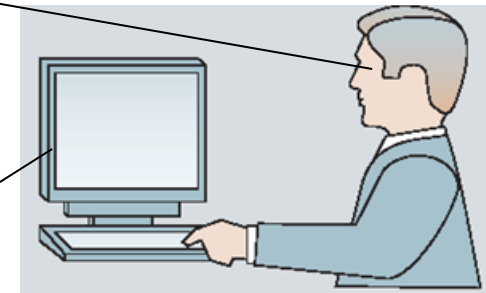


**To take a step further into Kidd lab,  
click on Drs. Ken & Judy Kidd to view the contents page and see where you would like to go.**

Done

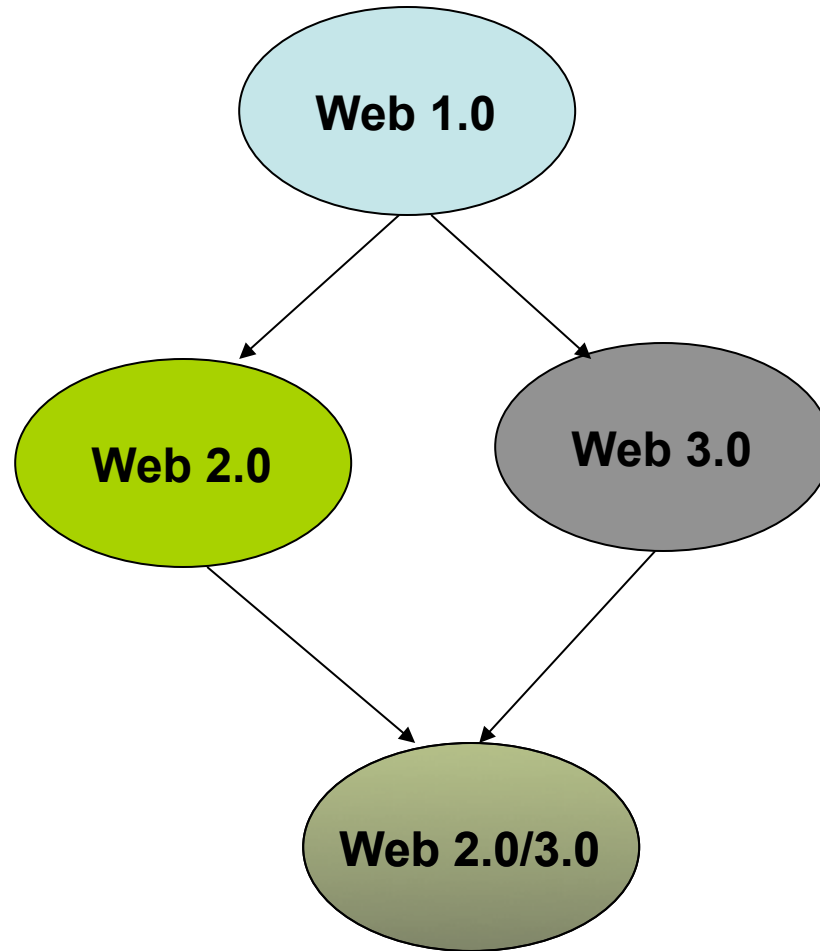
# Lack of Data Semantics

Type	Name	Synonym
Loci	Alcohol Dehydrogenase 1B (class I), beta polypeptide	ADH1B
Loci	Alcohol Dehydrogenase 1B (class I), beta polypeptide	ADH2
Loci	Solute carrier family 6 (neurotransmitter transporter, dopamine), member 3	ADHD
Loci	Alcohol Dehydrogenase 1C (class I), gamma polypeptide	ADH1C
Loci	Alcohol Dehydrogenase 1C (class I), gamma polypeptide	ADH3
Loci	Alcohol Dehydrogenase 7 (class IV), mu or sigma polypeptide	ADH-4
Loci	Alcohol Dehydrogenase 7 (class IV), mu or sigma polypeptide	ADH7



```
<html>
<body>
...
<table>
<tr>
<td><b>Type</b></td> <td><b>Name</b></td><td><b>Synonym</b></td>
</tr>
<tr>
<td>Loci</td> <td>Alcohol Dehydrogenase 1B (class I), beta polypeptide </td> <td>ADH1B </t>
</tr>
...
</table>
...
</body>
</html>
```

# Transforming Web 1.0 into Web 2.0 & Web 3.0



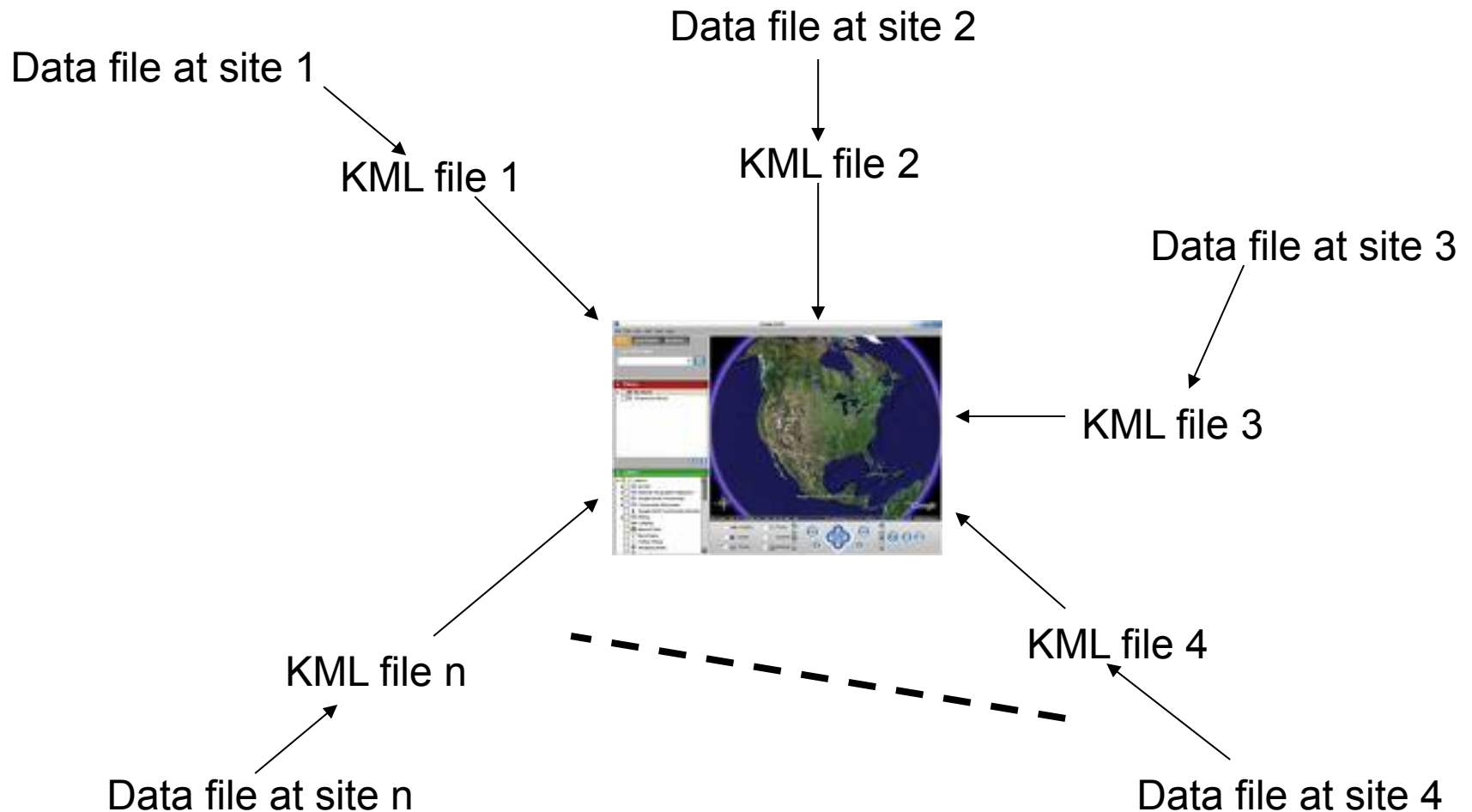
# Web 2.0

- It changes the way people communicate and share artifacts on the web (e.g., Flickr, youtube, facebook)
- Wiki, blog, RSS, folksonomy (social tagging)
- Multimedia rich (songs, images, videos, etc)
- Dynamic, interactive, responsive user interface (Ajax)
- XML-based data exchange format

# Mashup

- [Mashup \(music\)](#), the musical genre encompassing songs which consist entirely of parts of other songs
- [Mashup \(video\)](#), a video that is edited from more than one source to appear as one
- [Mashup \(digital\)](#), a digital media file containing any or all of text, graphics, audio, video, and animation, which recombines and modifies existing digital works to create a derivative work.
- [Mashup \(web application hybrid\)](#), a web application that combines data and/or functionality from more than one source

# XML(KML) - Based Geo Data Mashup





# Use of Yahoo! Pipes to convert tabular data into KML format

The screenshot shows the Yahoo! Pipes interface for a pipe named "State Cancer Profile". The workflow is as follows:

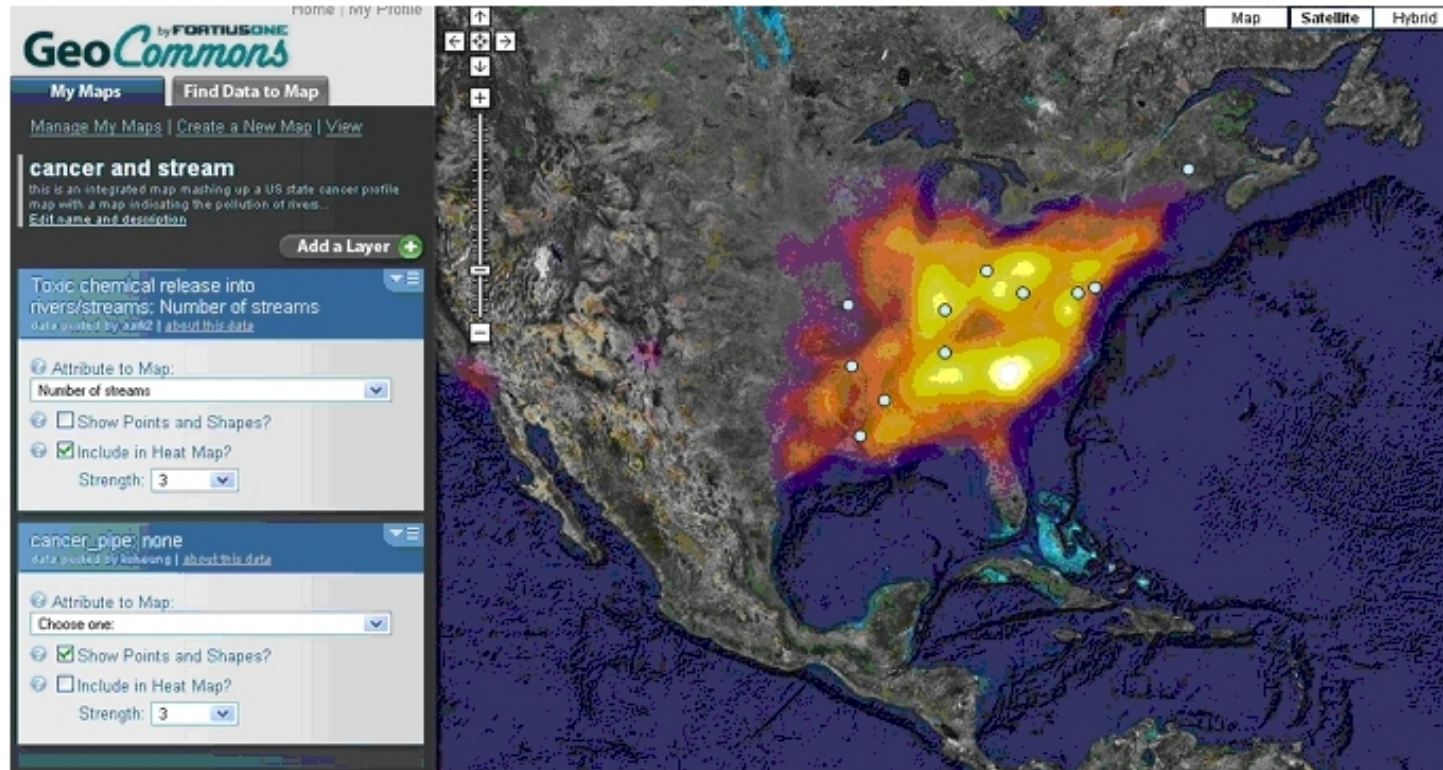
- Fetch CSV:** URL: `http://statecancerprofiles.cancer.gov`. Column separated by `,` or `and`. Skip the first `6` rows. Use the following column names: `state`, `annualDeathRate`, `lowerNinetyFivePer`, `upperNinetyFivePer`, `averageAnnualDea`, `ratePeriod`, `intervalRange`.
- Filter:** Permit items that match all of the following. Rule: `item.annualDeathRate` is greater than `200`.
- Rename:** Mappings: `item state` to `CopyAs title`, `item annualDeathRate` to `CopyAs description`.
- Location Extractor:** Extracts location information from the data.
- Pipe Output:** Outputs the data as a KML file.

The output is displayed as a map of the United States with red pins indicating the locations of the states with an annual death rate greater than 200. The map shows pins for states such as California, Texas, Florida, and others.

(a)

(b)

# GeoCommons: Mashup of Maps



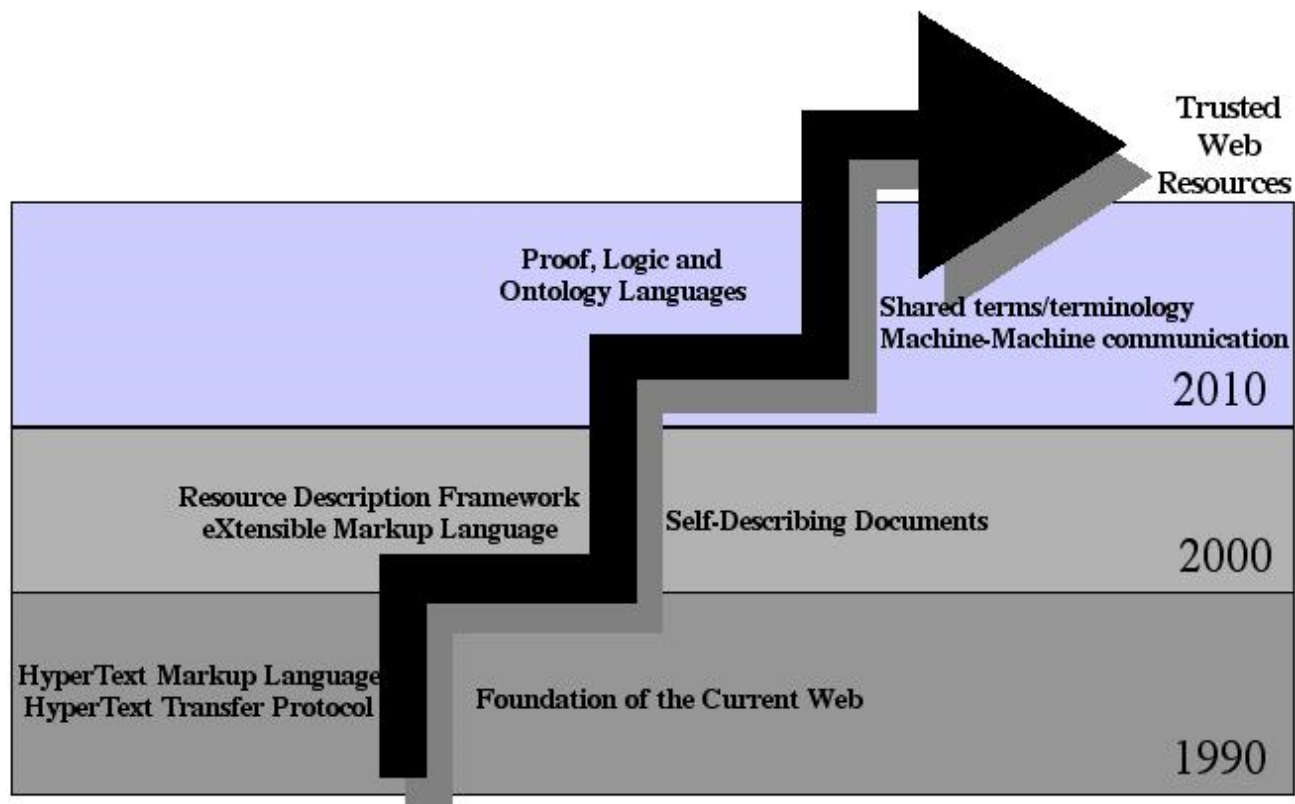
# What is an ontology?

- An ontology is a specification of a conceptualization
- It is a description of the concepts and their relationships that exist for a particular domain

# Web 3.0: Semantic Web

- The **Semantic Web** provides a common machine-readable ontology framework that allows **data** to be represented, shared and reused across application, enterprise, and community boundaries
  - The Semantic Web is a knowledge web of data
- The Semantic Web is about two things
  - It is about common formats for identification, representation, and integration of data drawn from diverse sources
  - It is also about languages for recording how the data relates to real world objects

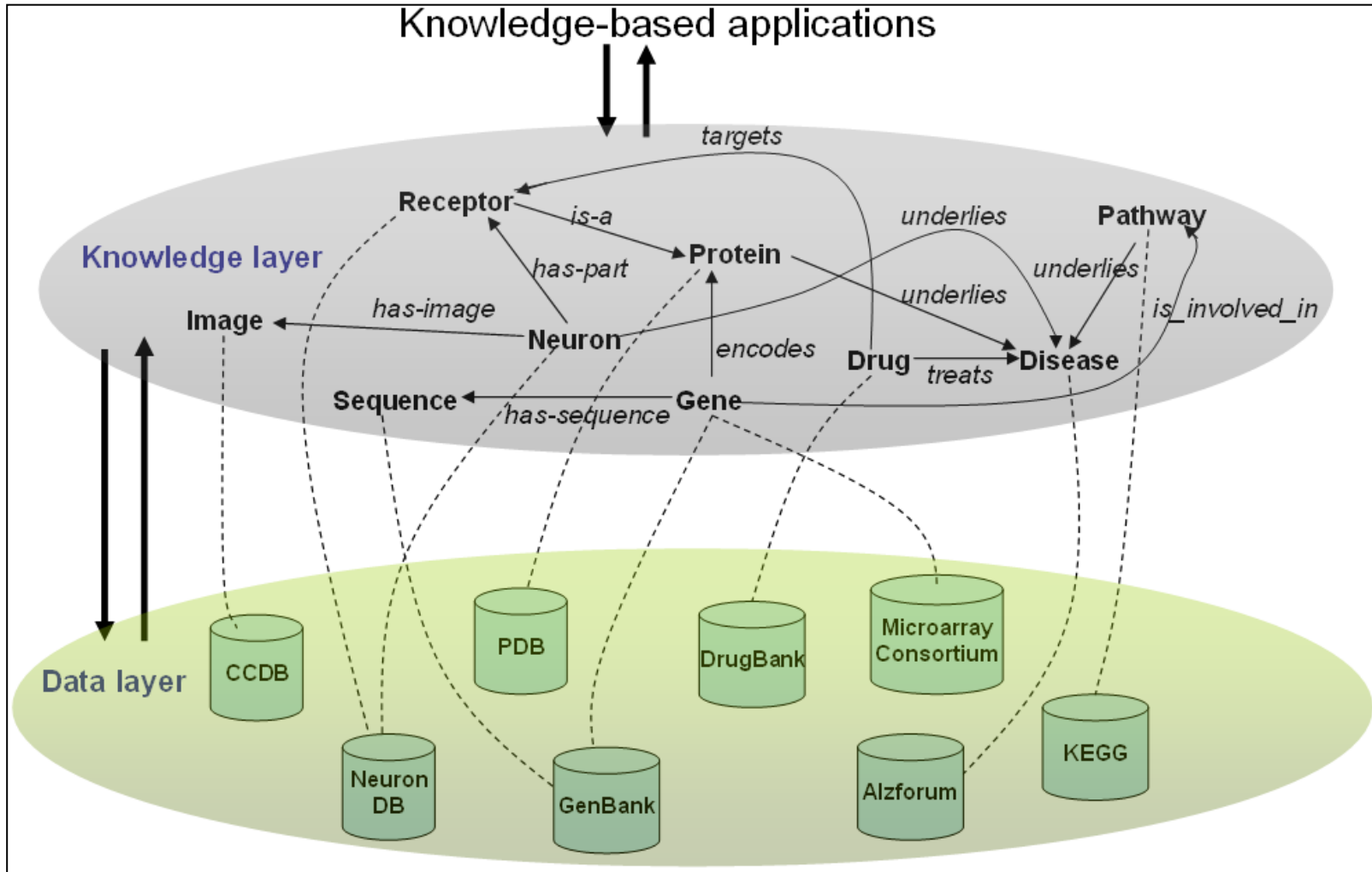
# Layers of the Semantic Web



Semantic Web = Brilliant Web



# Knowledge Web Data Integration



# Web 3.0: Semantic Web (Cont'd)

- Global identifying scheme (URI)
- Standard data modeling languages (RDF, RDFS, OWL)
- Standard query languages (SPARQL)
- Enabling tools/technologies (e.g., Protégé, Jena, triplestore, etc)



# Resource Description Framework (RDF)

- It is a standard data model (directed acyclic graph) for representing information (metadata) about resources in the World Wide Web
- In general, it can be used to represent information about “things” or “resources” that can be identified (using URI’s) on the Web
- It is intended to provide a simple way to make statements (descriptions) about Web resources

# Uniform Resource Identifiers (URIs)

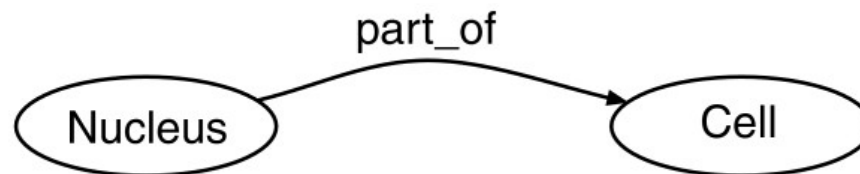
- A URI is a string of characters used to identify or name a resource on the Internet.
- URLs (Uniform Resource Locators) are a particular type of URI, used for resources that can be accessed on the WWW (e.g., web pages)
- In RDF, URIs typically look like “normal” URLs, often with fragment identifiers to point at specific parts of a document:
  - [http://www.semantic-systems-biology.org/SSB#CCO\\_B0000000](http://www.semantic-systems-biology.org/SSB#CCO_B0000000)  
(id for “core cell cycle protein” in Cell Cycle Ontology)

# RDF Triple/Graph

- The basic information unit in RDF is an RDF statement in the form of
  - (subject, property, object)
- Each RDF statement can be modeled as a graph comprising two nodes connected by a directed arc

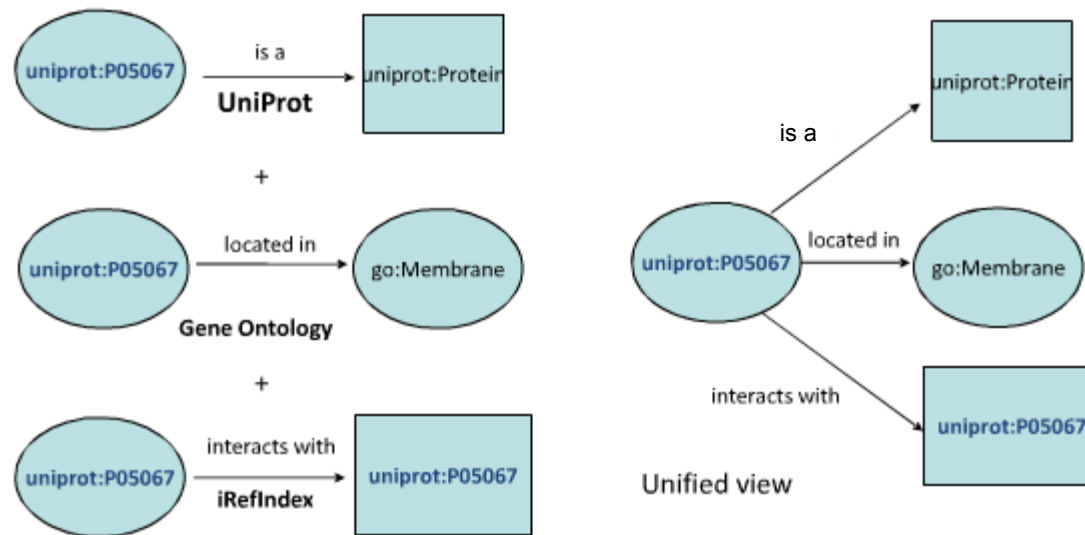


- A triple example

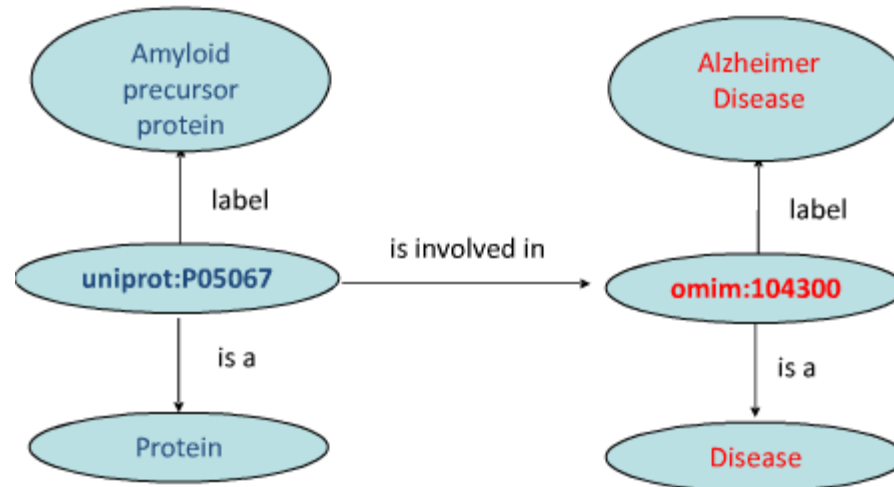


- A set of such triples can jointly form a directed labeled graph (DLG) that can in theory model a significant part of domain knowledge.
- An RDF graph can be represented in different formats (XML, Turtle, N3...)

# Linking data of the same type from multiple sources

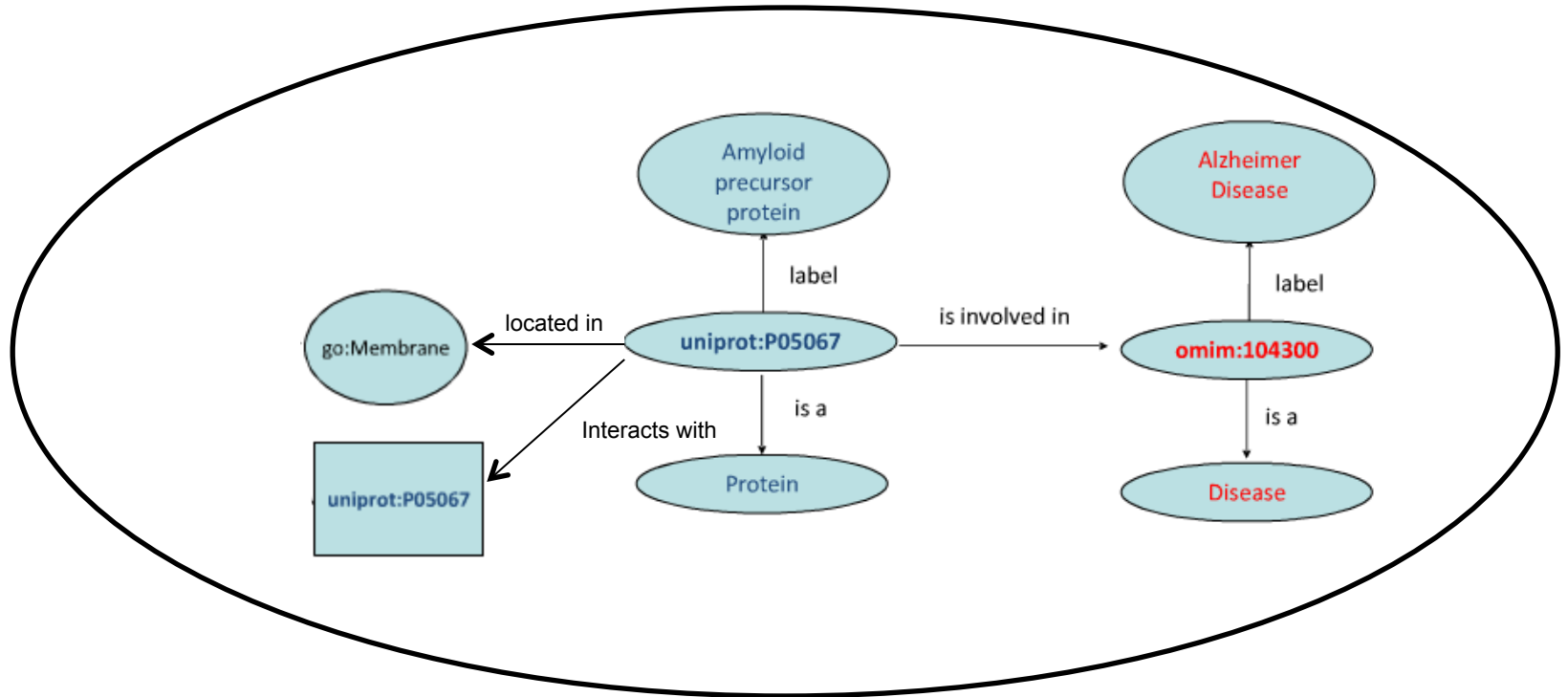


# Linking data across different types

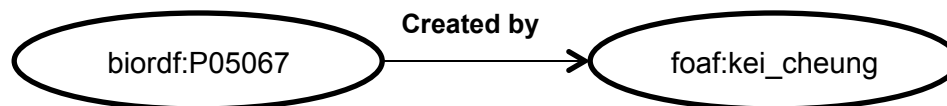


# Named Graph

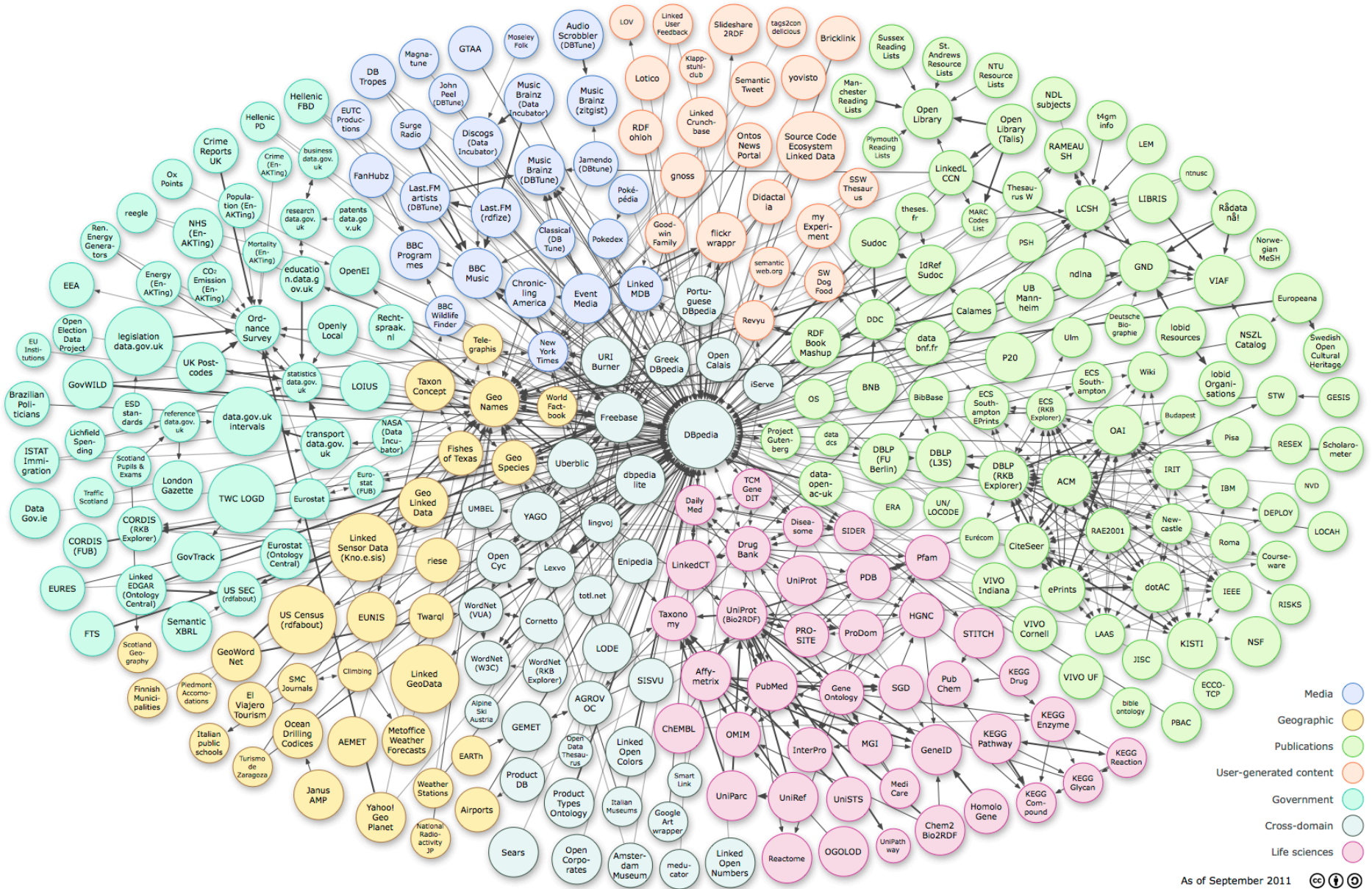
biordf:P05067



Meta Statement

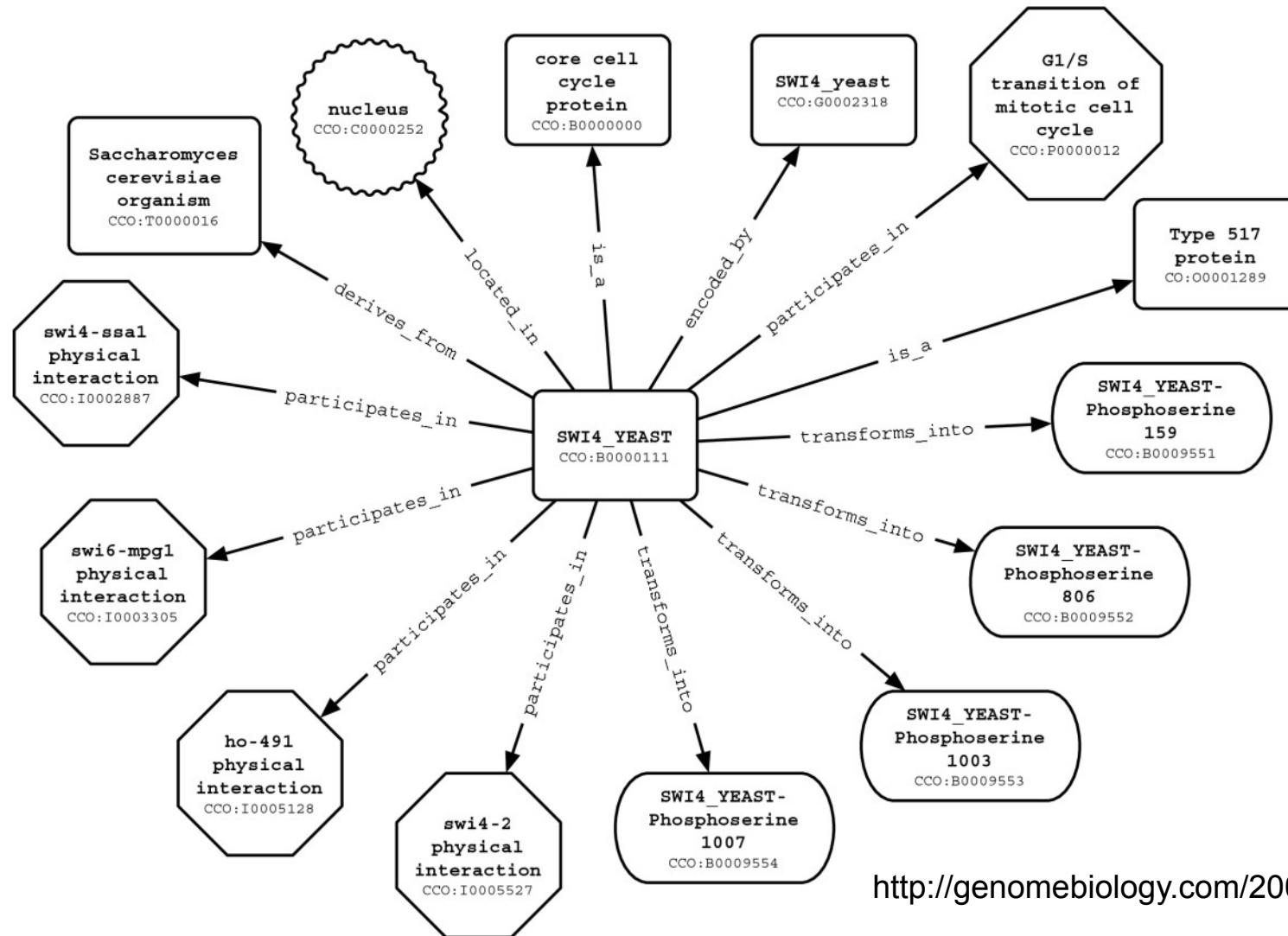


# Linked Data Cloud



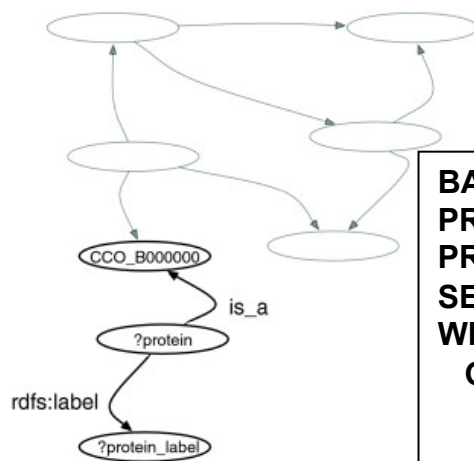
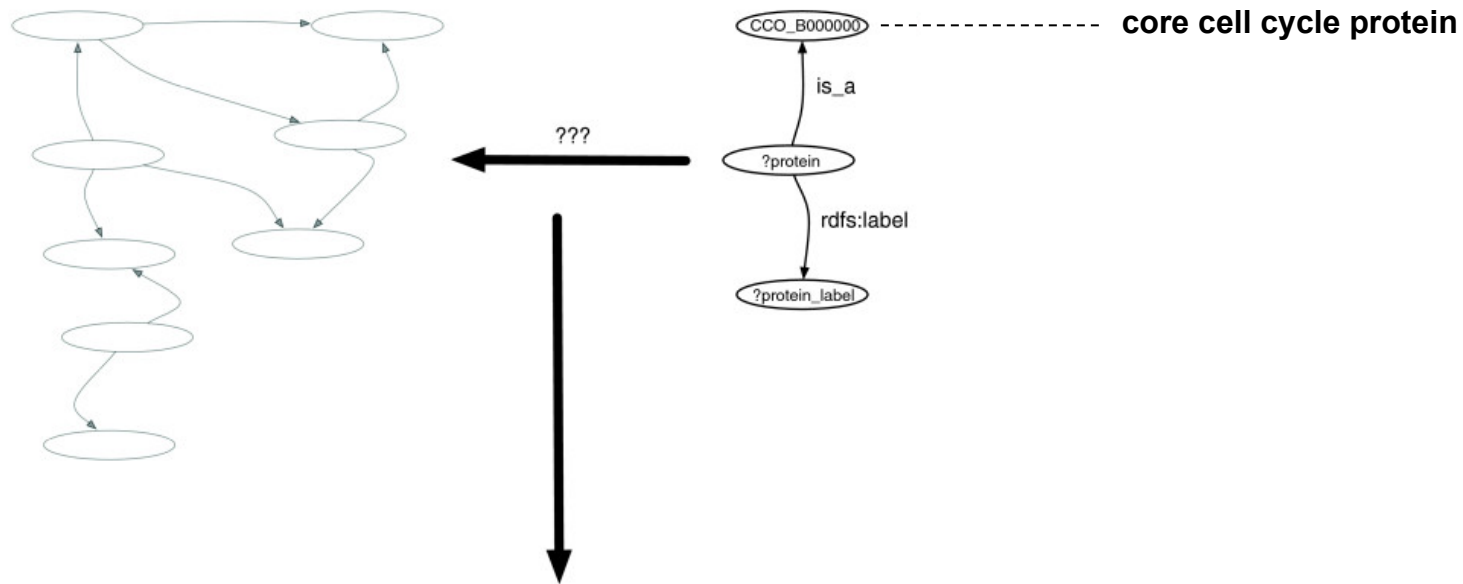
# Cell Cycle Ontology (CCO)

(Antezana et al, 2009, Genome Biology)





# RDF Graph Match (SPARQL)



```
BASE <http://www.semantic-systems-biology.org/webcite>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#webcite>
PREFIX ssb:<http://www.semantic-systems-biology.org/SSB#webcite>
SELECT ?protein_label
WHERE {
  GRAPH <cco_S_pombe> {
    ?protein ssb:is_a ssb:CCO_B000000.
    ?protein rdfs:label ?protein_label
  }
}
```

# RDF Schema (RDFS)

- RDF Schema terms:
  - Class
  - Property
  - type
  - subClassOf
  - range
  - Domain
- Example:
  - <DNASequence, type, Class>
  - <Promoter, subClassOf, DNASequence>
  - <Protein, type, Class>
  - <TranscriptionFactor, subClassOf, Protein>
  - <Bind, type, Property>
  - <Bind, domain, TranscriptionFactor>
  - <Bind, range, Promoter>

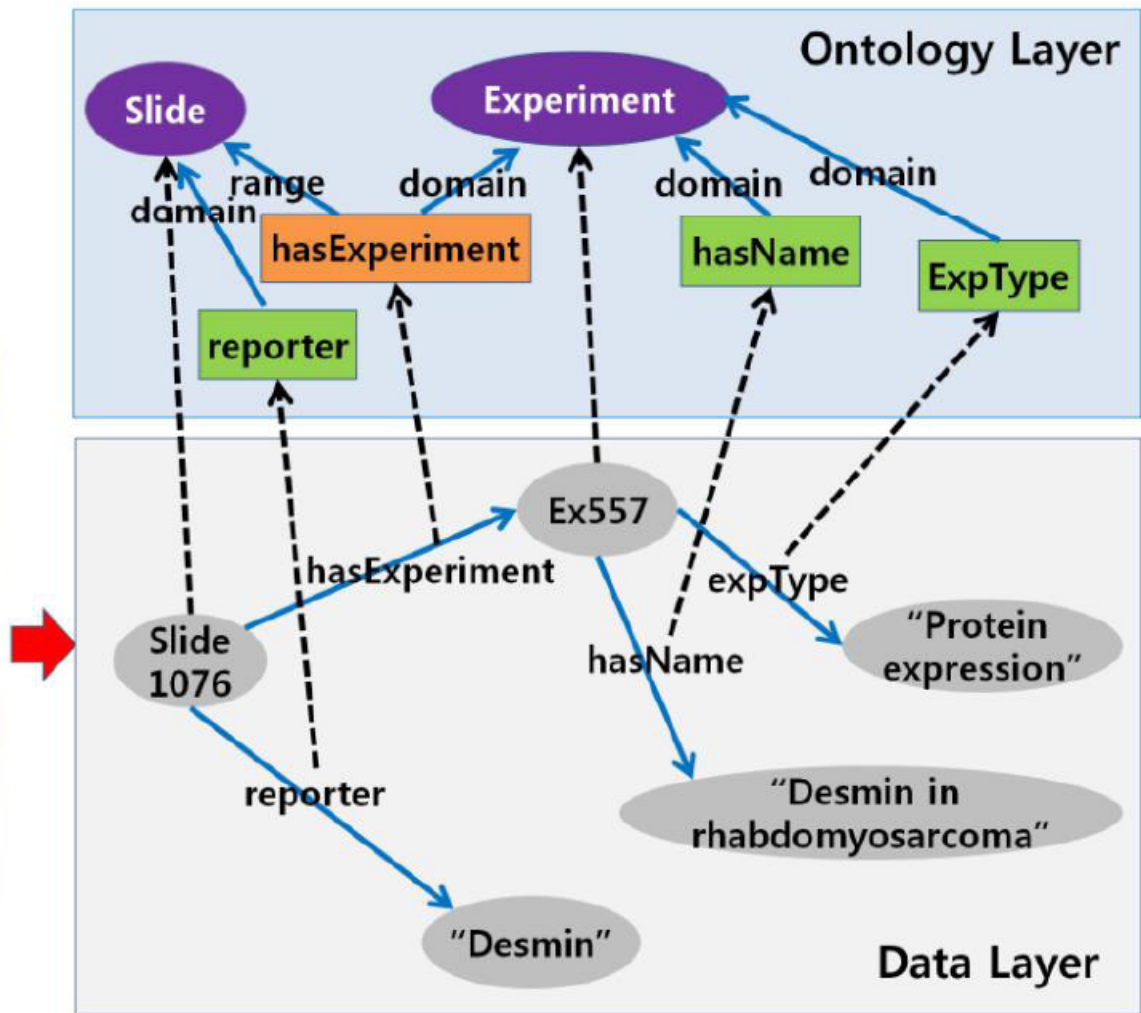
# Relational table -> RDF -> RDFS ontology

**Experiment table**

id	Name	ExpType
...	...	...
557	Desmin in rhabdomyosarcoma	Protein expression
...	....	...

**Slide table**

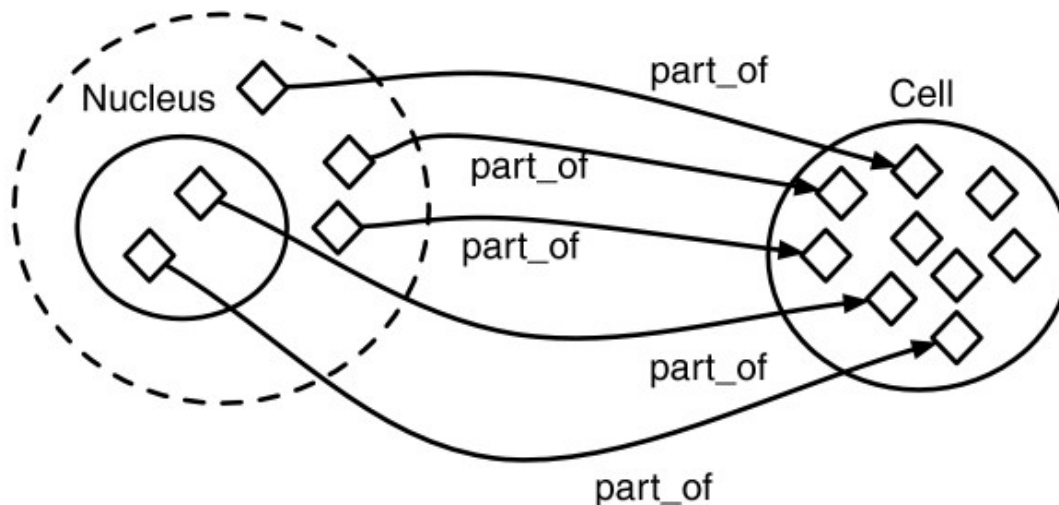
id	ExperimentId	Reporter
...	...	...
1076	557	Desmin
...	....	...



# Web Ontology Language (OWL)

- It is more semantically expressive than RDF and RDFS, but it is syntactically the same as RDF
  - Relationship constraints such as cardinality, sameAs, etc
- It has three species: OWL Lite, OWL DL, OWL Full

# OWL DL Representation (Subsumption)



```
:Nucleus  
a owl:Class ;  
rdfs:subClassOf  
  [ a owl:Restriction ;  
    owl:onProperty :part_of ;  
    owl:someValuesFrom :Cell  
  ]
```

**Necessary but not sufficient condition:** part of a nucleus is also part of a cell,  
but part of a cell is not necessarily part of a nucleus

# OWL Reasoning

- Which proteins participate in “mitosis”

```
:Protein  
  a owl:Class ;  
  rdfs:subClassOf  
    [ a owl:Restriction ;  
      owl:onProperty :participates_in ;  
      owl:someValuesFrom :Mitosis  
    ]
```

# Semantic Web Rule Language (SWRL = OWL + Rules)

**hasParent(?x1,?x2)  $\wedge$  hasBrother(?x2,?x3)  $\Rightarrow$  hasUncle(?x1,?x3)**

# SWRL Application (Pseudogene Ontology)

Rule	Antecedents	Consequents
R1	$\psi$ -gene $p$ has parent gene $g$ $p$ in segment $s$ $s$ has SD pair $d$ $d$ contains gene $g$	$p$ has_parent_in_duplicate_segment $d$
R2	$\psi$ -gene $p$ has parent in duplicate segment $d$ gene-count( $d$ ) > 0 pseudogene-count( $d$ ) > 0	$p$ has_not_only_parent_in_duplicate_segment $d$
R3	$\psi$ -gene $p$ has parent in duplicate segment $d$ gene-count( $d$ ) = 1 pseudogene-count( $d$ ) = 0	$p$ has_only_parent_in_duplicate_segment $d$
R4	$\psi$ -gene $p$ has only parent in duplicate segment $d$ Kimura-score( $p$ ) $\geq$ 0.4	$p$ has_quality <i>MaybeUnderPositiveSelection</i>
R5	$\psi$ -gene $p$ has only parent in duplicate segment $d$ Kimura-score( $p$ ) $\leq$ -0.4	$p$ has_quality <i>MaybeUnderNegativeSelection</i>
R6	$\psi$ -gene $p$ has only parent in duplicate segment $d$ Kimura-score( $p$ ) > -0.4 and < 0.4	$p$ has_quality <i>UnderNeutralSelection</i>
R7	$\psi$ -gene $p$ has not only parent in duplicate segment $d$ $p$ in segment $s$ $p$ is $pdist$ from start of $s$ $p$ has parent gene $g$ $g$ is $gdist$ from start of $d$ $\psi$ -gene $p2$ in segment $d$ $p2$ is $p2dist$ from start of $d$ $abs(p2dist - pdist) < abs(gdist - pdist)$ $abs(p2dist - pdist) < length(p)$	$p$ aligns_with $p2$



# Boimedical ontologies available in RDF/OWL format

- UniProt
- Gene Ontology
- NCI Metathesaurus
- Cell Ontology
- Sequence Ontology
- Protein Ontology
- These and many more ontologies are available in ontology repositories such as the NCBO BioPortal (<http://bioportal.bioontology.org/>)

# SW Enabling Technologies

- Ontology editor (e.g., protégé)
- Triple store (e.g., virtuoso)
- OWL reasoner (e.g., Pellet)
- SWRL reasoner (e.g., protégé plug-in)

# Collaborative Semantic Web Projects

- Protein motion (biosciences)
- Pseudogenes (biosciences)
- SenseLab (neurosciences)
- Pathways (translational medicine)
- Semantic Web for Health Care and Life Science Interest Group (chartered by W3C)
  - BioRDF task force

**The End**