@HCLS F2F, November 2007

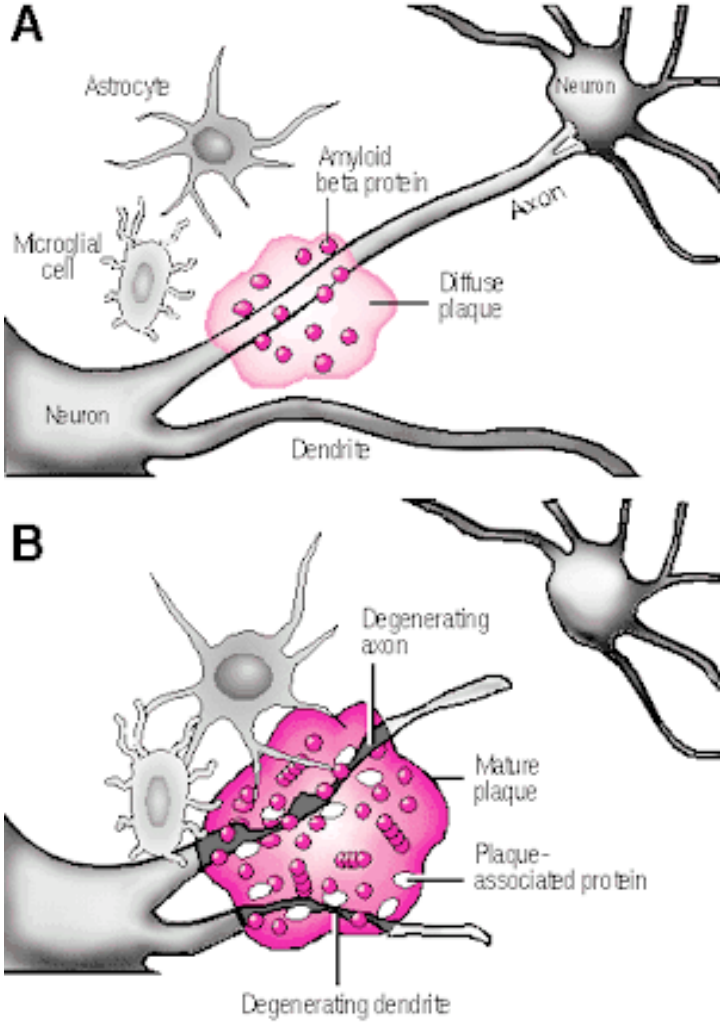
BioRDF: Understanding and enhancing the knowledge base

Alan Ruttenberg(Science Commons), speaking for a cast of thousands.

Three examples, then background, possibilities, and directions

- A simple SPARQL query
- Looking for Alzheimer's Disease targets
- Mashup: Google Maps/SPARQL/Allen Brain Atlas

A simple query: Biological processes in dendrites?



Alzheimer's disease is characterized by neural degeneration. Among other things, there is damage to dendrites and axons, parts of nerve cells.

What resources do we have available to learn more about biological processes in dendrites?

The Gene Ontology

The gene ontology names many biological processes and tells us which genes are known to be involved in those processes.

Let’s ask it what processes it knows about involving dendrites

Biological processes naming dendrites

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX go: <http://purl.org/obo/owl/GO#>
PREFIX obo: <http://www.geneontology.org/formats/oboInOwl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select  ?name  ?class ?definition
from <http://purl.org/commons/hcls/20070416>
where
{   graph <http://purl.org/commons/hcls/20070416/classrelations>
     {?class rdfs:subClassOf go:GO_0008150}
    ?class rdfs:label ?name.
    ?class obo:hasDefinition ?def.
    ?def rdfs:label ?definition
    filter(regex(?name,"[Dd]endrite"))
}
```

URI for Biological Process

From the “console”

result | request | response

name	class	definition
dendrite development	http://purl.org/obo/owl/GO#GO_0016358	The process whose specific outcome is the progression of the dendrite over time, from its formation to the mature structure. A dendrite is a freely branching protoplasmic process of a nerve cell.
dendrite regeneration	http://purl.org/obo/owl/GO#GO_0031104	The regrowth of dendrites following their loss or damage.
dendrite morphogenesis	http://purl.org/obo/owl/GO#GO_0048813	The process by which the anatomical structures of dendrite are generated and organized. Morphogenesis pertains to the creation of form. A dendrite is a freely branching protoplasmic process of a nerve cell.
regulation of dendrite morphogenesis	http://purl.org/obo/owl/GO#GO_0048814	Any process that modulates the frequency, rate or extent of dendrite morphogenesis.
regulation of dendrite development	http://purl.org/obo/owl/GO#GO_0050773	Any process that modulates the frequency, rate or extent of dendrite development.
negative regulation of dendrite morphogenesis	http://purl.org/obo/owl/GO#GO_0050774	Any process that stops, prevents or reduces the frequency, rate or extent of dendrite morphogenesis.
positive regulation of dendrite morphogenesis	http://purl.org/obo/owl/GO#GO_0050775	Any process that activates or increases the frequency, rate or extent of dendrite morphogenesis.

But the answers are also available by a "GET"

```
/sparql/?query=PREFIX%20owl%3A%20%3Chttp%3A%2F%2Fwww.w3.org
%2F2002%2F07%2Fowl%23%3E%0APREFIX%20go%3A%20%3Chttp%3A%2F%2
Fpurl.org%2Fobo%2Fowl%2FGO%23%3E%0APREFIX%20obo%3A%20%3Chtt
p%3A%2F%2Fwww.geneontology.org%2Fformats%2FoboInOwl%23%3E%0
APREFIX%20rdfs%3A%20%3Chttp%3A%2F%2Fwww.w3.org%2F2000%2F01%
2Frdf-
schema%23%3E%0A%0Aselect%20%20%3Fname%20%20%3Fclass%20%3Fde
finition%0Afrom%20%3Chttp%3A%2F%2Fpurl.org%2Fcommons%2Fhcls
%2F20070416%3E%0Awhere%0A%7B%20%20%20graph%20%3Chttp%3A%2F%
2Fpurl.org%2Fcommons%2Fhcls%2F20070416%2Fclassrelations%3E%
0A%20%20%20%20%20%7B%3Fclass%20rdfs%3AsubClassOf%20go%3AGO_
0008150%7D%0A%20%20%20%20%3Fclass%20rdfs%3Alabel%20%3Fname.
%0A%20%20%20%20%3Fclass%20obo%3AhasDefinition%20%3Fdef.%0A%
20%20%20%20%3Fdef%20rdfs%3Alabel%20%3Fdefinition%20%0A%20%2
0%20%20filter(regex(%3Fname%2C%22%5BDd%5Dendrite%22))%0A%7D
%0A&format=&maxrows=50
```

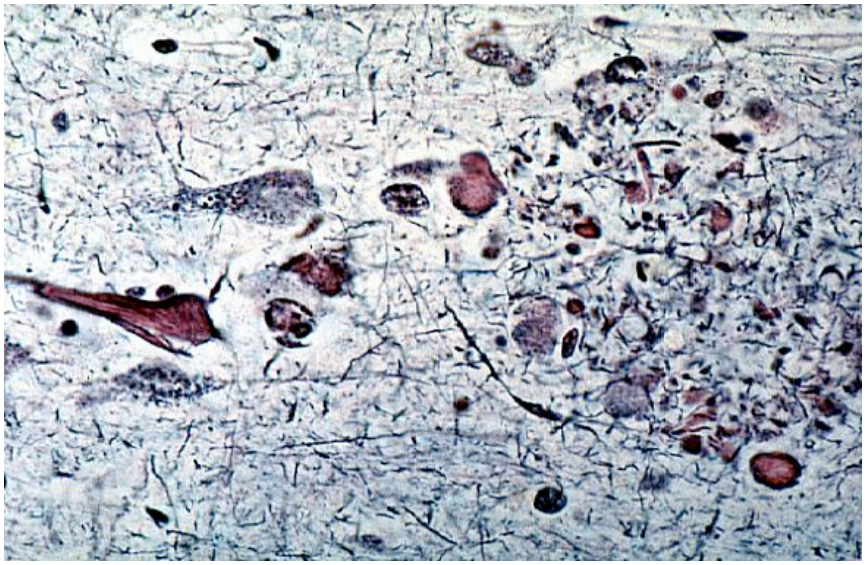
So someone, somewhere else, can build something better

Looking for Alzheimer Disease targets

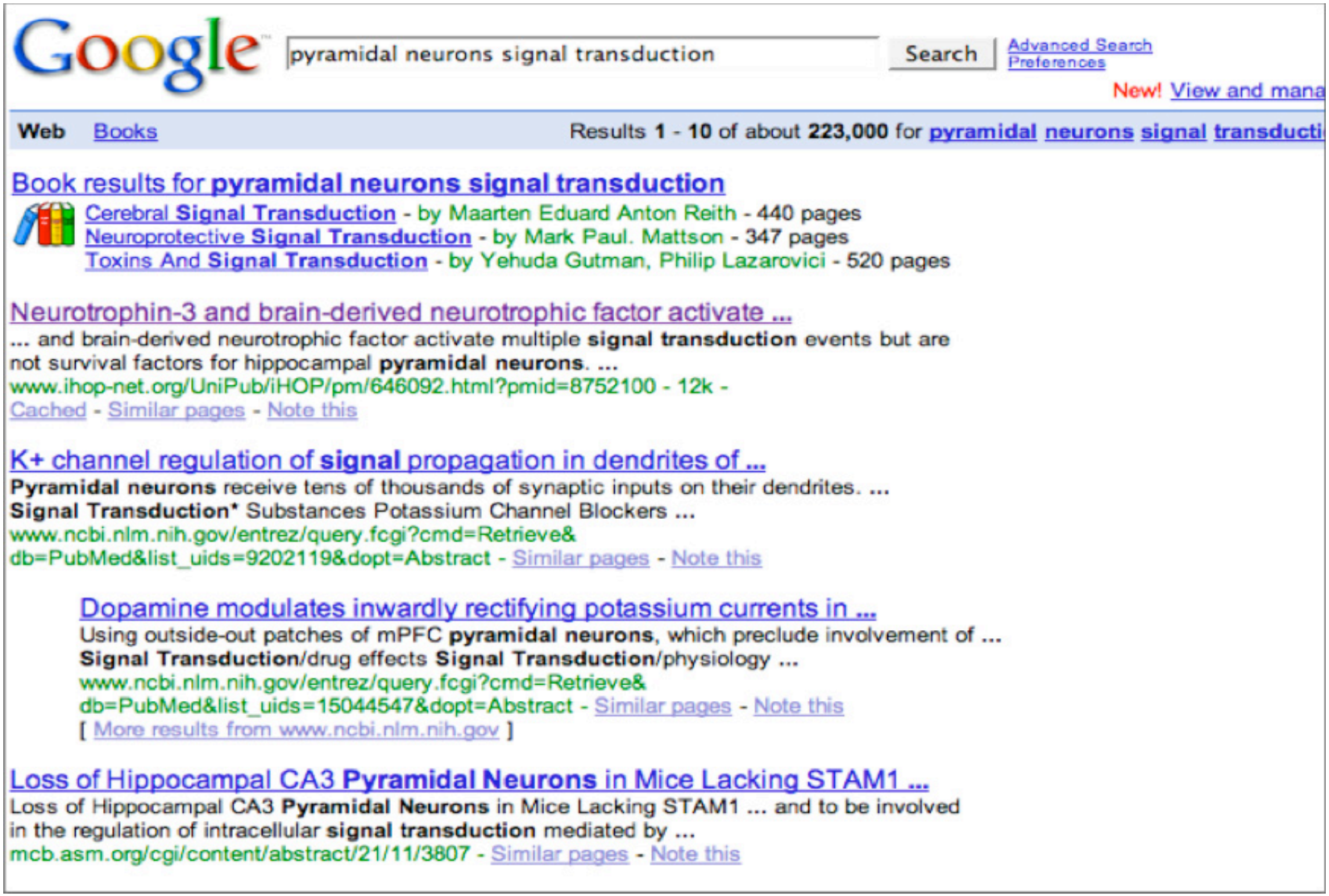
Signal transduction pathways are considered to be rich in “druggable” targets - proteins that might respond to chemical therapy

CA1 Pyramidal Neurons are known to be particularly damaged in Alzheimer’s disease.

Casting a wide net, can we find candidate genes known to be involved in signal transduction and active in Pyramidal Neurons?



Google: 223,000 results



Google pyramidal neurons signal transduction Search Advanced Search Preferences

New! View and mana

Web Books Results 1 - 10 of about 223,000 for pyramidal neurons signal transducti

Book results for **pyramidal neurons signal transduction**

Cerebral **Signal Transduction** - by Maarten Eduard Anton Reith - 440 pages
Neuroprotective **Signal Transduction** - by Mark Paul. Mattson - 347 pages
Toxins And **Signal Transduction** - by Yehuda Gutman, Philip Lazarovici - 520 pages

Neurotrophin-3 and brain-derived neurotrophic factor activate ...
... and brain-derived neurotrophic factor activate multiple **signal transduction** events but are not survival factors for hippocampal **pyramidal neurons**. ...
www.ihop-net.org/UniPub/iHOP/pm/646092.html?pmid=8752100 - 12k -
Cached - Similar pages - Note this

K+ channel regulation of **signal** propagation in dendrites of ...
Pyramidal neurons receive tens of thousands of synaptic inputs on their dendrites. ...
Signal Transduction* Substances Potassium Channel Blockers ...
www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=9202119&dopt=Abstract - Similar pages - Note this

Dopamine modulates inwardly rectifying potassium currents in ...
Using outside-out patches of mPFC **pyramidal neurons**, which preclude involvement of ...
Signal Transduction/drug effects **Signal Transduction**/physiology ...
www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=15044547&dopt=Abstract - Similar pages - Note this
[More results from www.ncbi.nlm.nih.gov]

Loss of Hippocampal CA3 **Pyramidal Neurons** in Mice Lacking STAM1 ...
Loss of Hippocampal CA3 **Pyramidal Neurons** in Mice Lacking STAM1 ... and to be involved in the regulation of intracellular **signal transduction** mediated by ...
mcb.asm.org/cgi/content/abstract/21/11/3807 - Similar pages - Note this

A SPARQL query spanning 4 sources

```
prefix go: <http://purl.org/obo/owl/GO#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix mesh: <http://purl.org/commons/record/mesh/>
prefix sc: <http://purl.org/science/owl/sciencecommons/>
prefix ro: <http://www.obofoundry.org/ro/ro.owl#>

select ?genename ?processname
where
{  graph <http://purl.org/commons/hcls/pubmesh>
    { ?paper ?p mesh:D017966 .
      ?article sc:identified_by_pmid ?paper.
      ?gene sc:describes_gene_or_gene_product_mentioned_by ?article.
    }
  graph <http://purl.org/commons/hcls/goa>
    { ?protein rdfs:subClassOf ?res.
      ?res owl:onProperty ro:has_function.
      ?res owl:someValuesFrom ?res2.
      ?res2 owl:onProperty ro:realized_as.
      ?res2 owl:someValuesFrom ?process.
   graph <http://purl.org/commons/hcls/20070416/classrelations>
    {{?process <http://purl.org/obo/owl/obo#part_of> go:GO_0007166}
      union
     {?process rdfs:subClassOf go:GO_0007166 }}
      ?protein rdfs:subClassOf ?parent.
      ?parent owl:equivalentClass ?res3.
      ?res3 owl:hasValue ?gene.
     }
  graph <http://purl.org/commons/hcls/gene>
    { ?gene rdfs:label ?genename }
  graph <http://purl.org/commons/hcls/20070416>
    { ?process rdfs:label ?processname}
}
```

Mesh: Pyramidal Neurons

↓

Pubmed: Journal Articles

↓

Entrez Gene: Genes

↓

GO: Signal Transduction

Inference required

Results

Many of the genes are indeed related to Alzheimer's Disease through gamma secretase (presenilin) activity

DRD1, 1812	adenylate cyclase activation
ADRB2, 154	adenylate cyclase activation
ADRB2, 154	arrestin mediated desensitization of G-protein coupled receptor protein signaling pathway
DRD1IP, 50632	dopamine receptor signaling pathway
DRD1, 1812	dopamine receptor, adenylate cyclase activating pathway
DRD2, 1813	dopamine receptor, adenylate cyclase inhibiting pathway
GRM7, 2917	G-protein coupled receptor protein signaling pathway
GNG3, 2785	G-protein coupled receptor protein signaling pathway
GNG12, 55970	G-protein coupled receptor protein signaling pathway
DRD2, 1813	G-protein coupled receptor protein signaling pathway
ADRB2, 154	G-protein coupled receptor protein signaling pathway
CALM3, 808	G-protein coupled receptor protein signaling pathway
HTR2A, 3356	G-protein coupled receptor protein signaling pathway
DRD1, 1812	G-protein signaling, coupled to cyclic nucleotide second messenger
SSTR5, 6755	G-protein signaling, coupled to cyclic nucleotide second messenger
MTNR1A, 4543	G-protein signaling, coupled to cyclic nucleotide second messenger
CNR2, 1269	G-protein signaling, coupled to cyclic nucleotide second messenger
HTR6, 3362	G-protein signaling, coupled to cyclic nucleotide second messenger
GRIK2, 2898	glutamate signaling pathway
GRIN1, 2902	glutamate signaling pathway
GRIN2A, 2903	glutamate signaling pathway
GRIN2B, 2904	glutamate signaling pathway
ADAM10, 102	integrin-mediated signaling pathway
GRM7, 2917	negative regulation of adenylate cyclase activity
LRP1, 4035	negative regulation of Wnt receptor signaling pathway
ADAM10, 102	Notch receptor processing
ASCL1, 429	Notch signaling pathway
HTR2A, 3356	serotonin receptor signaling pathway
ADRB2, 154	transmembrane receptor protein tyrosine kinase activation (dimerization)
PTPRG, 5793	transmembrane receptor protein tyrosine kinase signaling pathway
EPHA4, 2043	transmembrane receptor protein tyrosine kinase signaling pathway
NRTN, 4902	transmembrane receptor protein tyrosine kinase signaling pathway
CTNND1, 1500	Wnt receptor signaling pathway

Some questions you might care about answering

For what neurological disorders are cell lines available?

For Parkinsons disease, what tissue and cell lines are available?

Give me information on the receptors and channels expressed in cortical neurons

What chemical agents can be used visualizing the nervous system?

A question I was asked

Create a system that will let us prioritize an expected 2000 siRNA hits according to whether there is chemical matter for studying them, e.g. validated antibodies, since we can only follow up on 600.

We know how to use Semantic Web technology to answer these kinds of questions (but there is no free lunch)

The Allen Brain Atlas

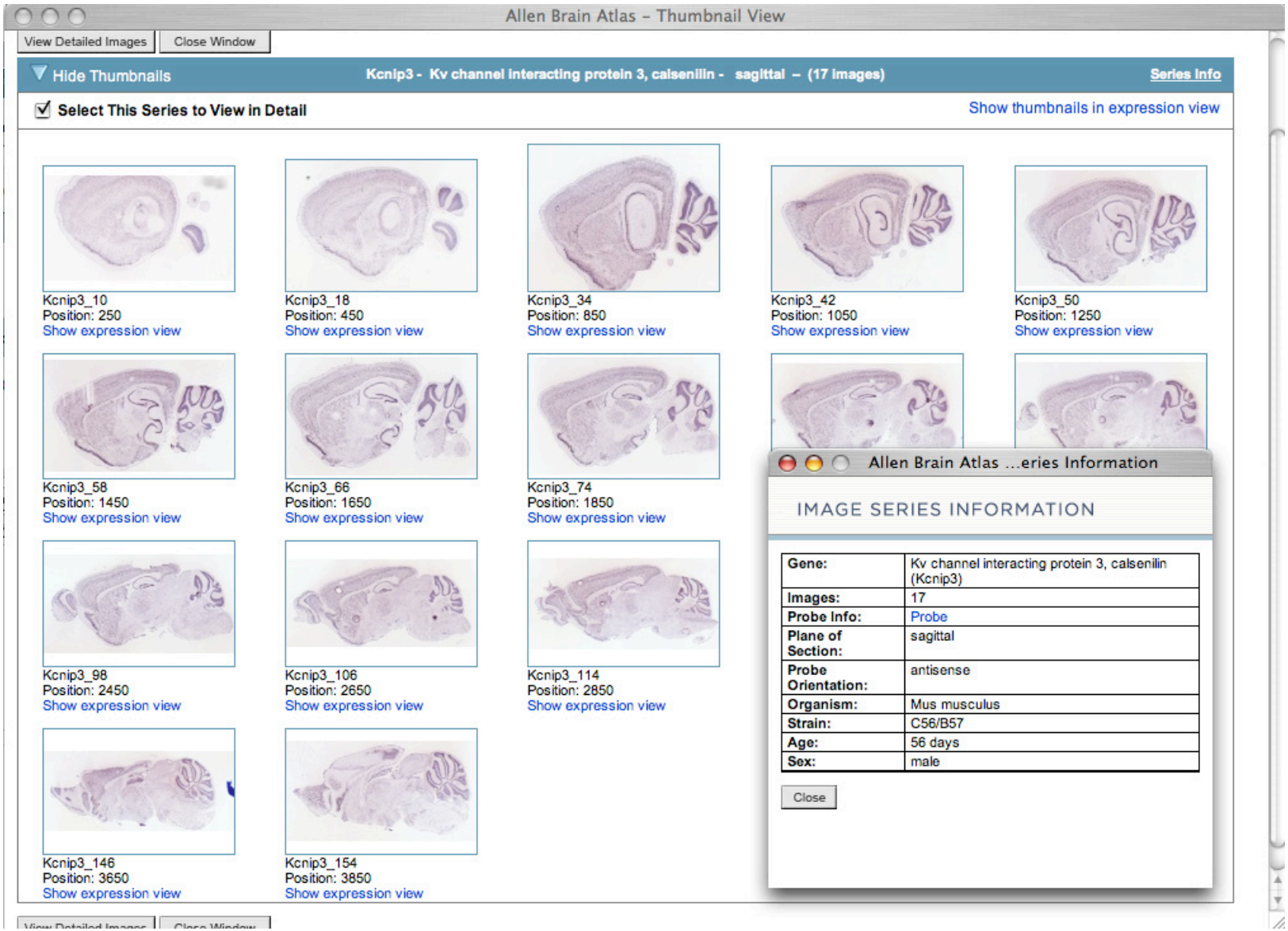
A remarkable scientific achievement. Mouse brains cut into thin slices and stained for the presence of gene expression.

20,000 genes, 400000 images at high resolution.

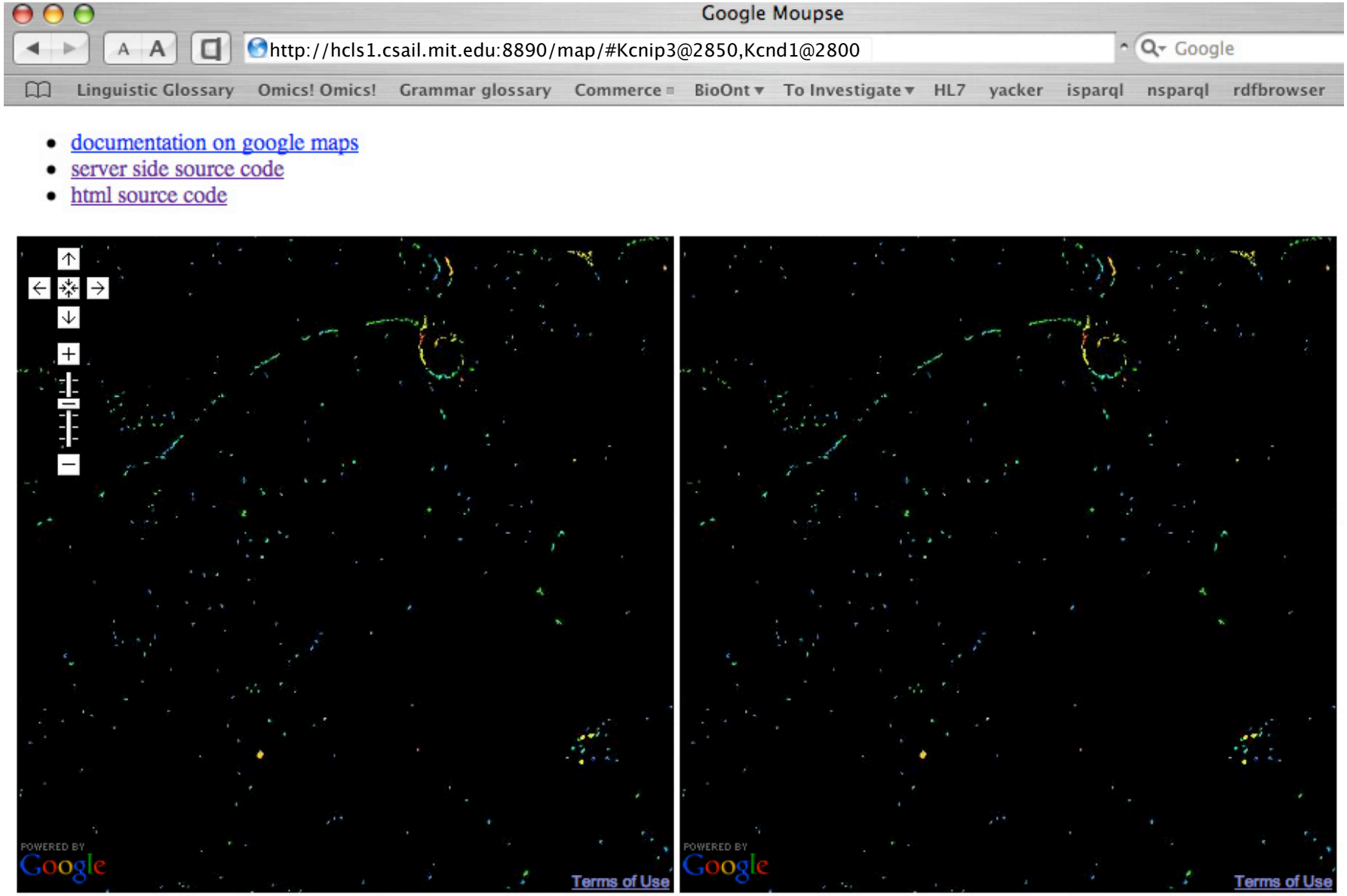
Currently available only through an HTML interface.

Scrape 80K web pages to extract the information, convert to RDF

The Allen Brain Atlas



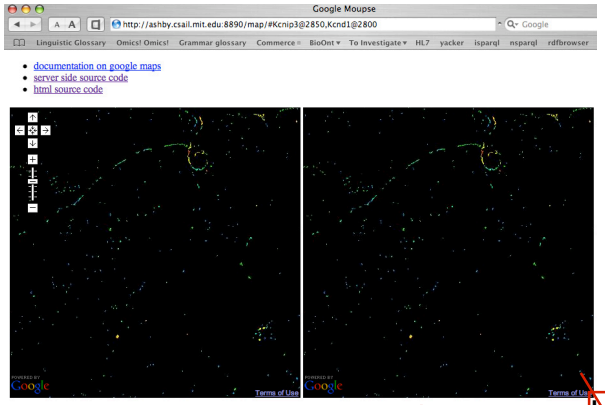
Google Maps/SPARQL/Allen Brain Atlas



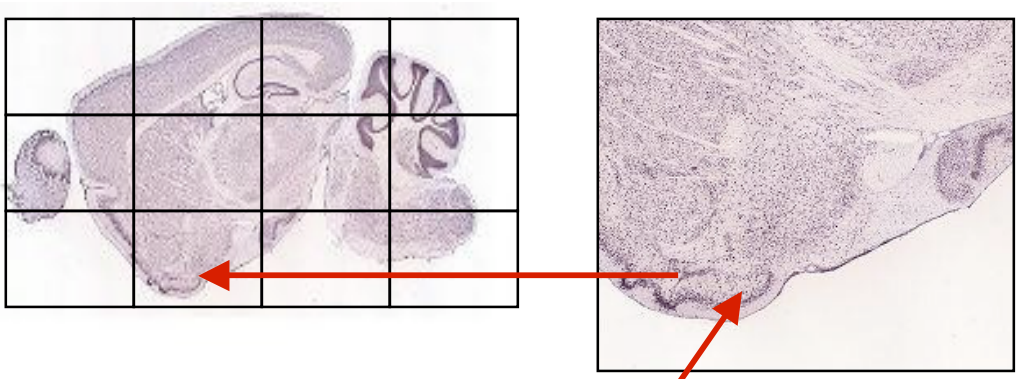
How it works (standing on the shoulders of giants)

http://hcls1/map/#Kcnip3@2850,Kcnd1@2800

Javascript



Allen Brain Institute Servers



http://www.brainmap.org://….0205032816_B.aff/TileGroup3/1-0-1.jpg

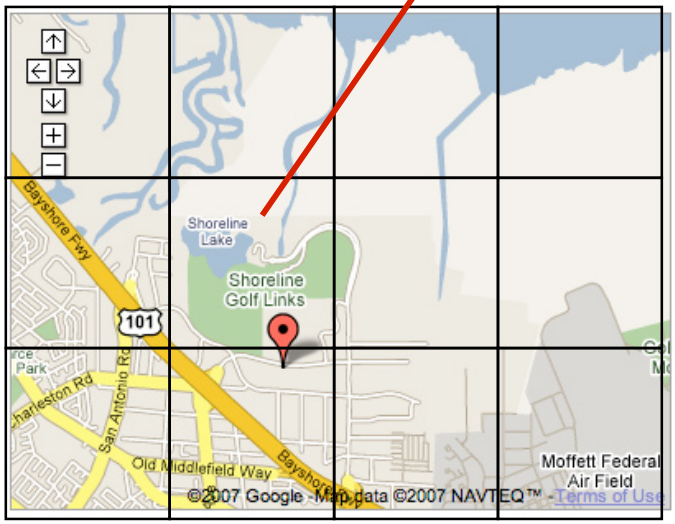
SPARQL
AJAX

Query

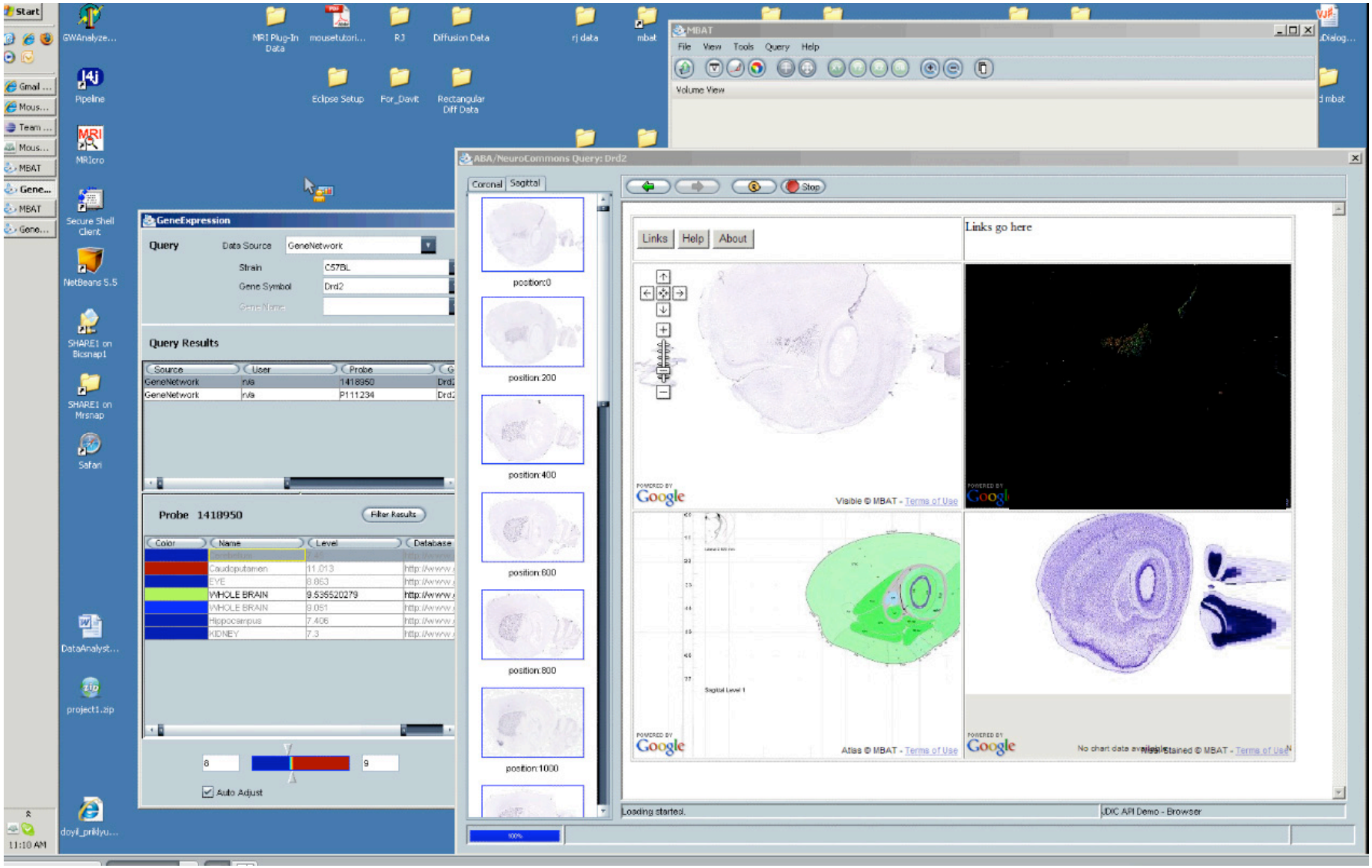
URL

HCLS/Neurocommons Servers

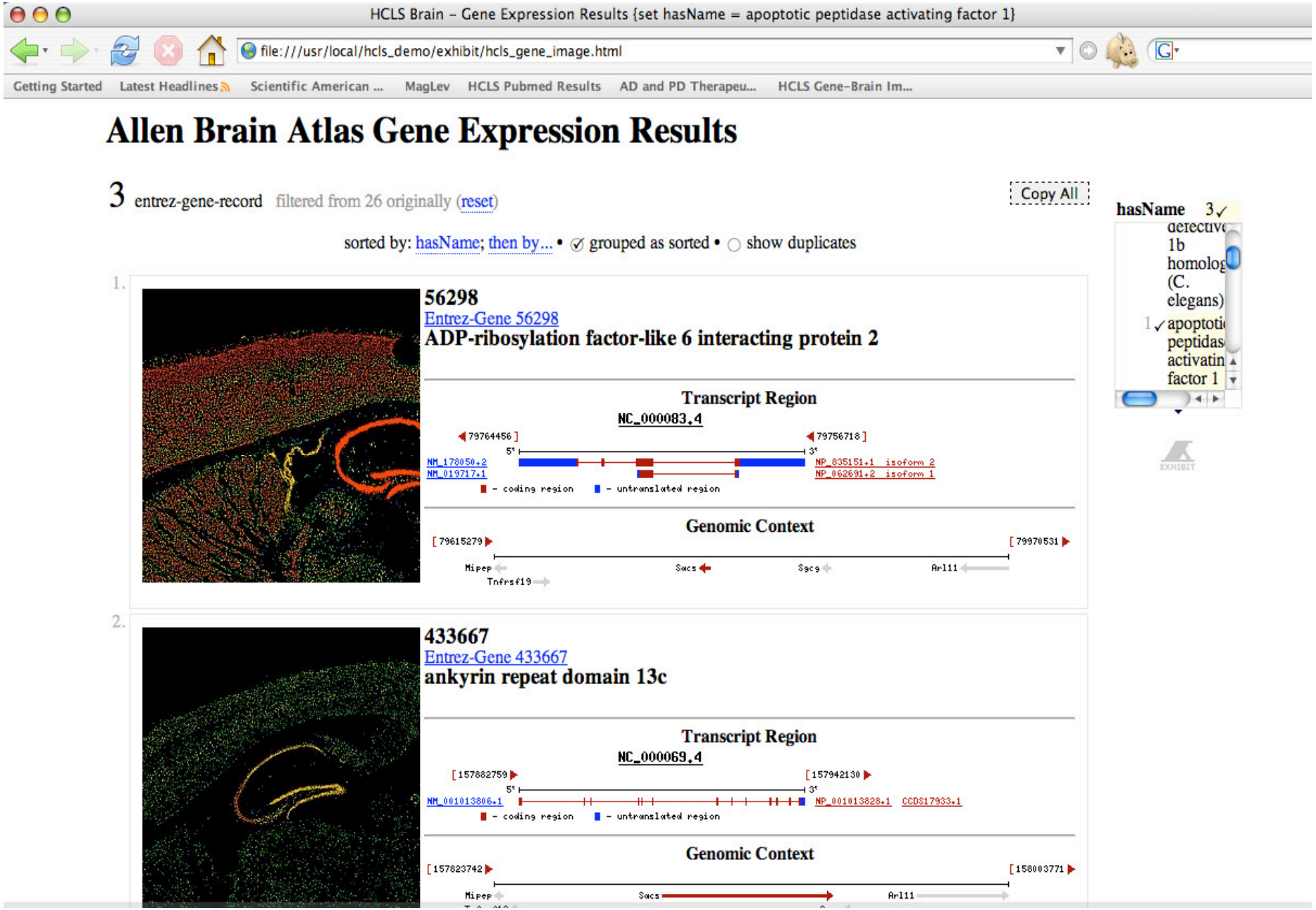
Google
Maps
API



BIRN can “view source”, use our code in MBAT, just like people learning by using others’ html



But you can also use Exhibit to visualize



Data Integration for question answering at Web Scale

Two choices

- Easier to publish: Many schema, burden is on user to learn and query each, or on aggregators to combine
- Harder to publish: Fewer schema, burden is on producer/collaborator to learn to speak common language

We've aimed at common language, but we're not there yet

"If we look at the RDF/OWL datasets that are currently part of the 'HCLS demo' we can see that their structures are quite heterogeneous. Every data source is structured in a very unique way, so that someone writing a query spanning several data sources needs a deep understanding of each data source to make it work."

Matthias Samwald

Challenges and goals for the HCLS Semantic Web community in the next years

A strategy for consensus on representation for science

Define what terms mean by relating them (tracing them) to elements in reality. Or say when you are not.

Have a theory of what an instance corresponds to. (denotes)

Classes are defined in terms of instances of specified types (bundle of shared properties)

Figure out how to document and organize all this knowledge in a way that can be managed in a distributed manner.

The product of this effort is an Ontology

Instances (1)

"Objects" (*particulars, independent continuants*)

- A mouse
- A molecule
- A record in pubmed
- A book
- A syringe of blood
- A goose
- A flock of geese
- A microarray chip
- A person
- A printout of a journal article

"fully present at every time when it exists"

Instances (2)

"properties" (*dependent continuants*)

- The ability of a specific molecule instance to act as a catalyst (a function)
- What the MGH institutional review board can do (a role)
- One person's internal temperature (a quality) (changes over time)

"fully present at every time when it exists"

Instances(3)

Processes

- The mouse running across the floor
- The IRB deciding whether a specific study should be approved (realizing their role)
- A reaction in which a caspase cleaves a single protein (executing its function)
- One patient visiting a doctor for 1/2 an hour

“Takes place(unfolds) over a period of time”

Relations between instances

Hyatt second floor **above** Hyatt first floor

Jonathan **located_in** Hyatt

Alan **has_quality** {temperature of 98.6 farenheight} now

Eric **has_role** W3C Liason for HCLS

Classes

Those entities that are like in some way

Mostly expressible as the relationships that their instances have to other instances

A **water molecule** is a **molecule** that **has** 3 **parts** - **2 hydrogen atoms and 1 oxygen atom**.

A nurse is a **person** that **has role NurseRole**

- ***ALL instances of nurse are an instance of person that has_role SOME instance of NurseRole***

Words mash up functions and objects

Ligand

Neurotransmitter

Hormone

Peptide

Looking for peptides?

Normalized representations dissect words

PeptideReceptorLigand - **A peptide** that has a function which makes it able to bind to a receptor

PeptideNeurotransmitter - **A peptide** expressed in a neuron that has a function which makes it able to regulate another neuron

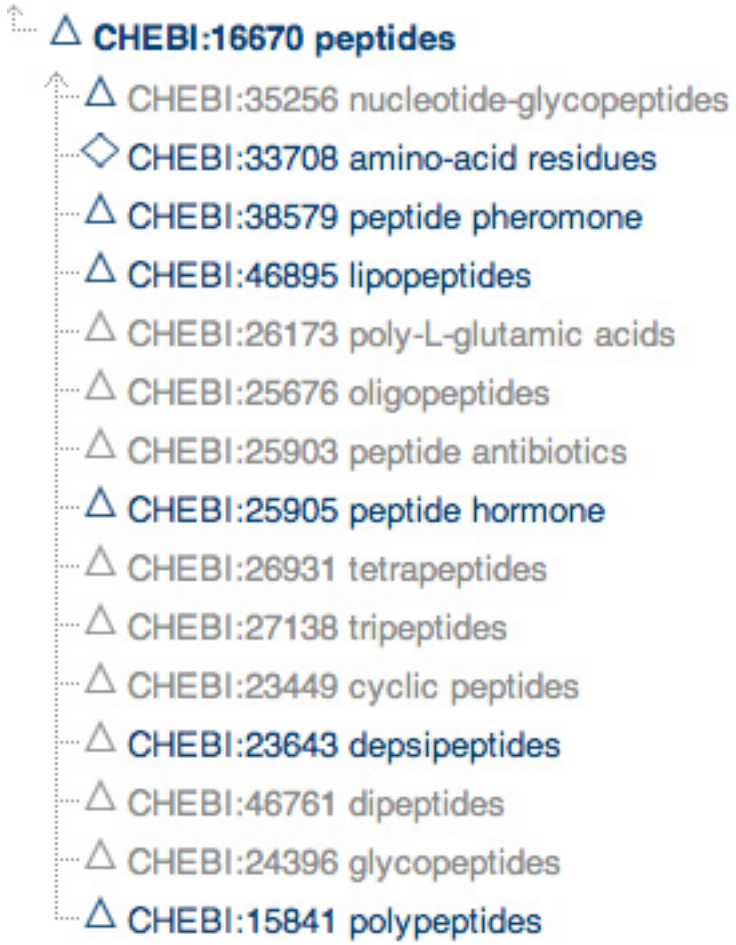
PeptideHormone - **A peptide** that produced in one organ and having an regulatory effect in another.

Peptide - A “short” polymer of amino acids

Looking for peptides?

Peptides from CHEBI

Chemical Entities of Biological Interest



Hormone Activity from GO Molecular Function

- all : all [439998]
  - GO:0003674 : molecular_function [286402]
    - GO:0005488 : binding [86988]
      - GO:0005515 : protein binding [42794]
        - GO:0005102 : receptor binding [3784]
          - **GO:0005179 : hormone activity [604]**
            - GO:0017045 : adrenocorticotropin-releasing hormone activity [2]
            - GO:0017044 : alpha-melanocyte stimulating hormone activity [0]
            - GO:0046659 : digestive hormone activity [0]
            - GO:0008613 : diuretic hormone activity [12]
            - GO:0016913 : follicle-stimulating hormone activity [4]
            - GO:0016608 : growth hormone-releasing hormone activity [16]
            - GO:0005183 : luteinizing hormone-releasing factor activity [22]
            - GO:0030354 : melanin-concentrating hormone activity [4]
            - GO:0016085 : myoinhibitory hormone activity [2]
            - GO:0016084 : myostimulatory hormone activity [2]
            - GO:0005184 : neuropeptide hormone activity [222]
              - **GO:0016087 : ecdysiostatic hormone activity [2]**
            - GO:0016521 : pituitary adenylate cyclase activating polypeptide activity [0]
            - GO:0008437 : thyrotropin-releasing hormone activity [6]

hormone activity

Term Information

Accession GO:0005179

Ontology molecular function

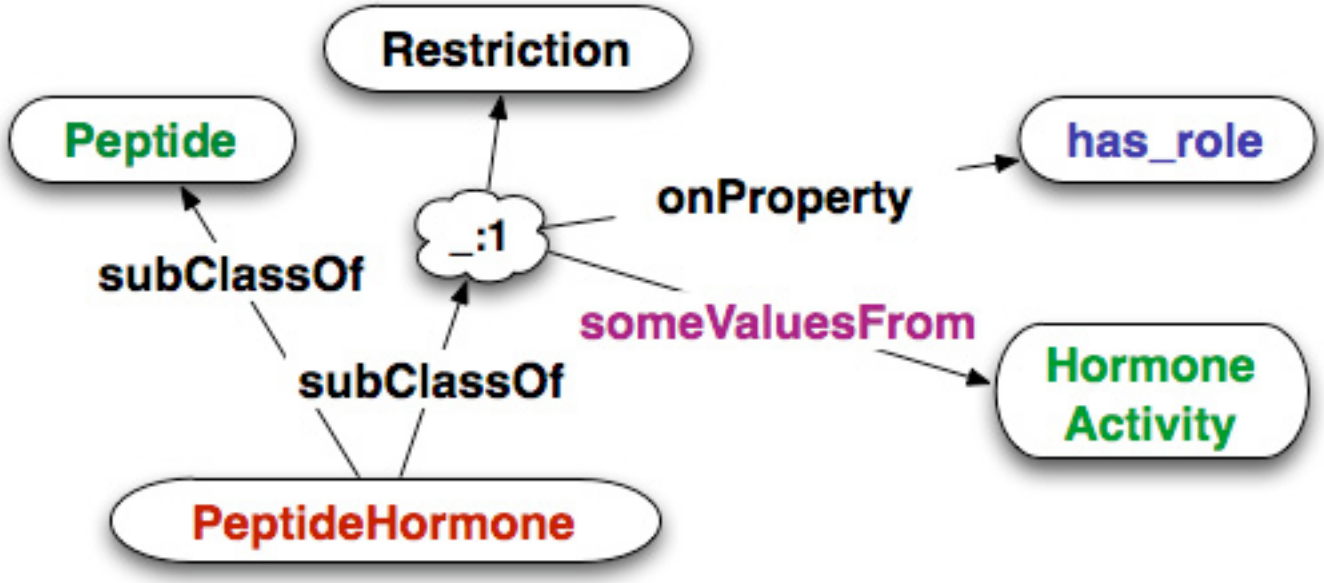
Synonyms
narrow: cAMP generating peptide activity
narrow: glycopeptide hormone
narrow: lipopeptide hormone
narrow: peptide hormone

Definition The action characteristic of a hormone, any substance formed in very small amounts in one specialized organ or group of cells and carried (sometimes in the bloodstream) to another organ or group of cells in the same organism, upon which it has a specific regulatory action. The term was originally applied to agents with a stimulatory physiological action in vertebrate animals (as opposed to a chalone, which has a depressant action). Usage is now extended to regulatory compounds in lower animals and plants, and to synthetic substances having comparable effects. [source: GOC:mah, ISBN:0198506732]

Comment None

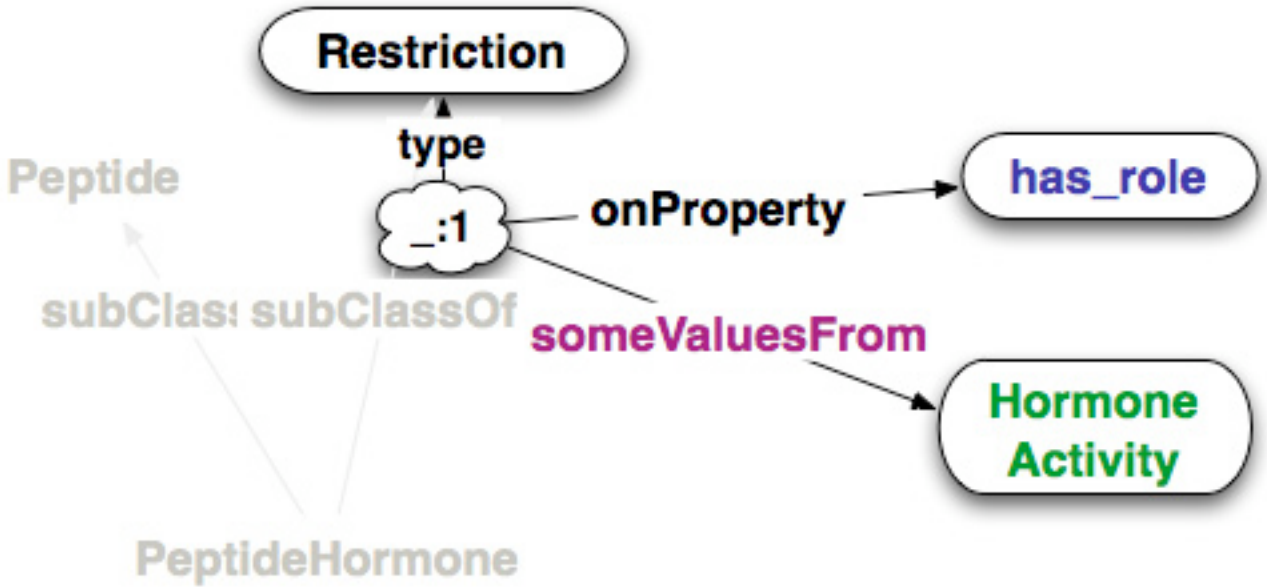
Towards RDF/OWL(1)

ALL* instances of *PeptideHormone* are an instance of *Peptide* that *has_role* *SOME* instance of *HormoneActivity

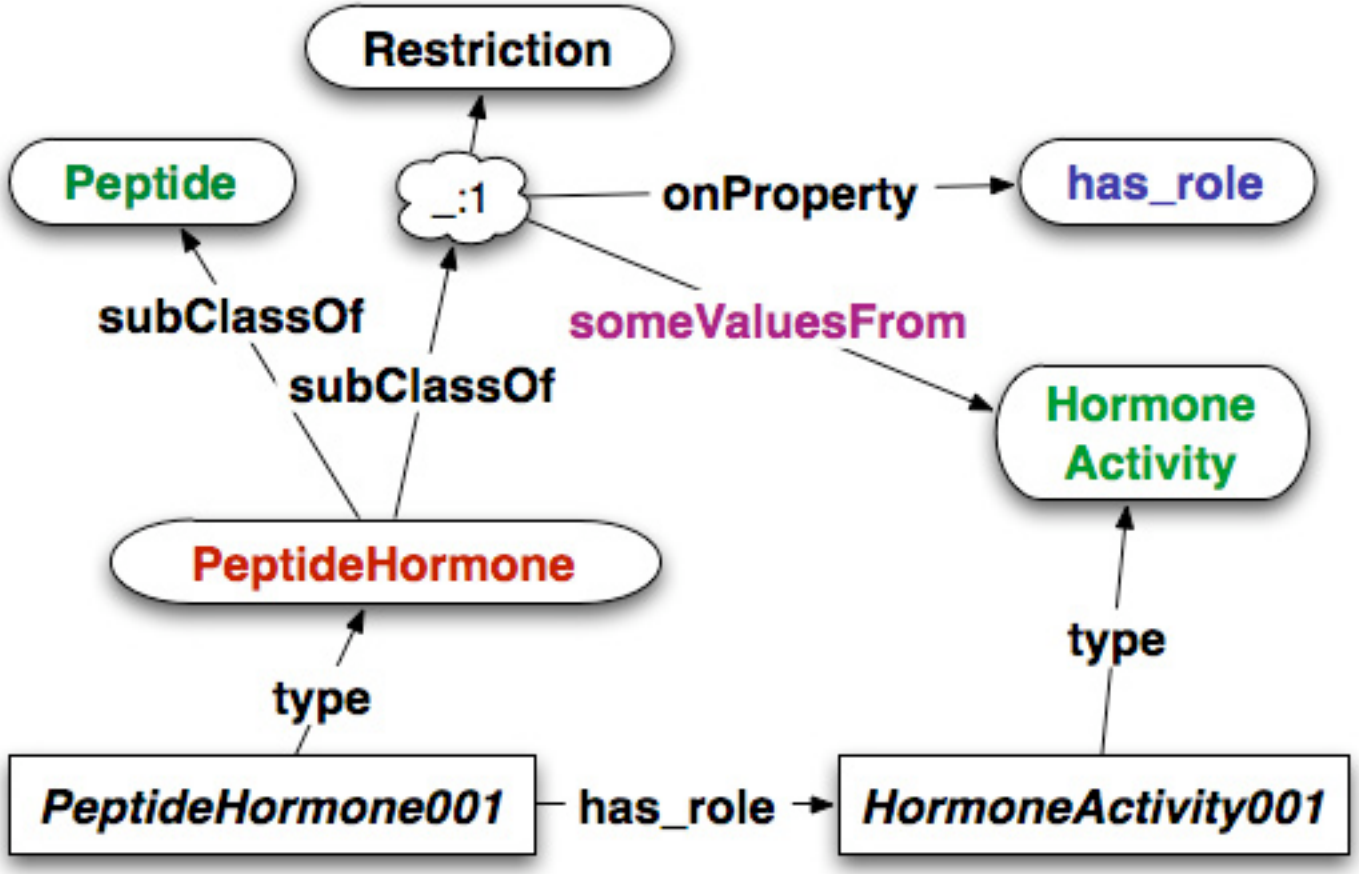


Towards RDF/OWL(3)

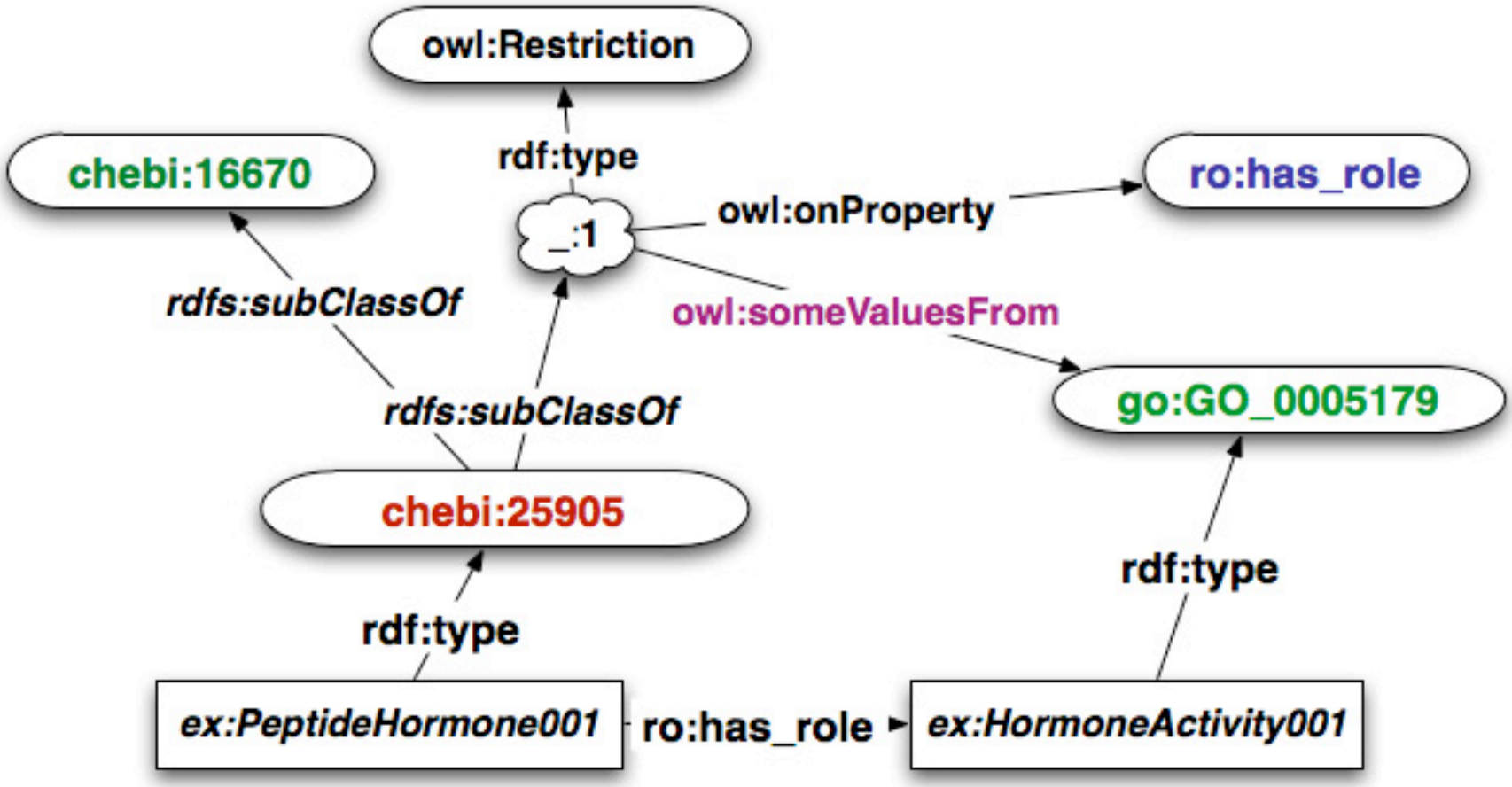
ALL instan **Class of things** *rmone are an instance of*
Peptide ***tha*** ***instance of HormoneActivity***



Towards RDF/OWL(3) - Instances

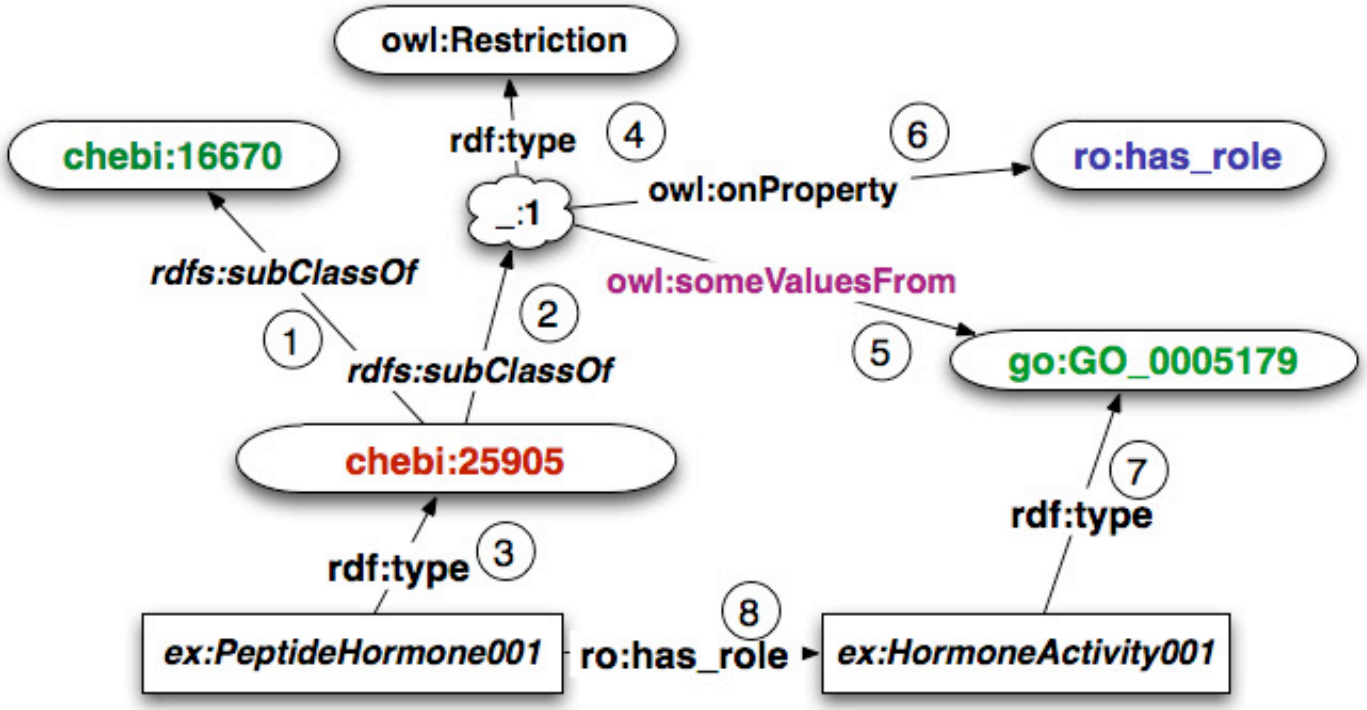


Towards RDF/OWL(4) URIs



chebi:25905 = <http://purl.org/obo/owl/CHEBI#CHEBI_25905>

Towards OWL(5) : triples



1. chebi:25905 rdfs:subClassOf chebi:16670.
2. chebi:25905 rdfs:subClassOf _:1.
6. :_1 owl:onProperty ro:hasRole.
5. :_1 owl:someValuesFrom go:GO_00179.

…

SPARQLing: Put ?variables where you are looking for matches

1 chebi:25905 rdfs:subClassOf chebi:16670.

2 chebi:25905 rdfs:subClassOf _:1.

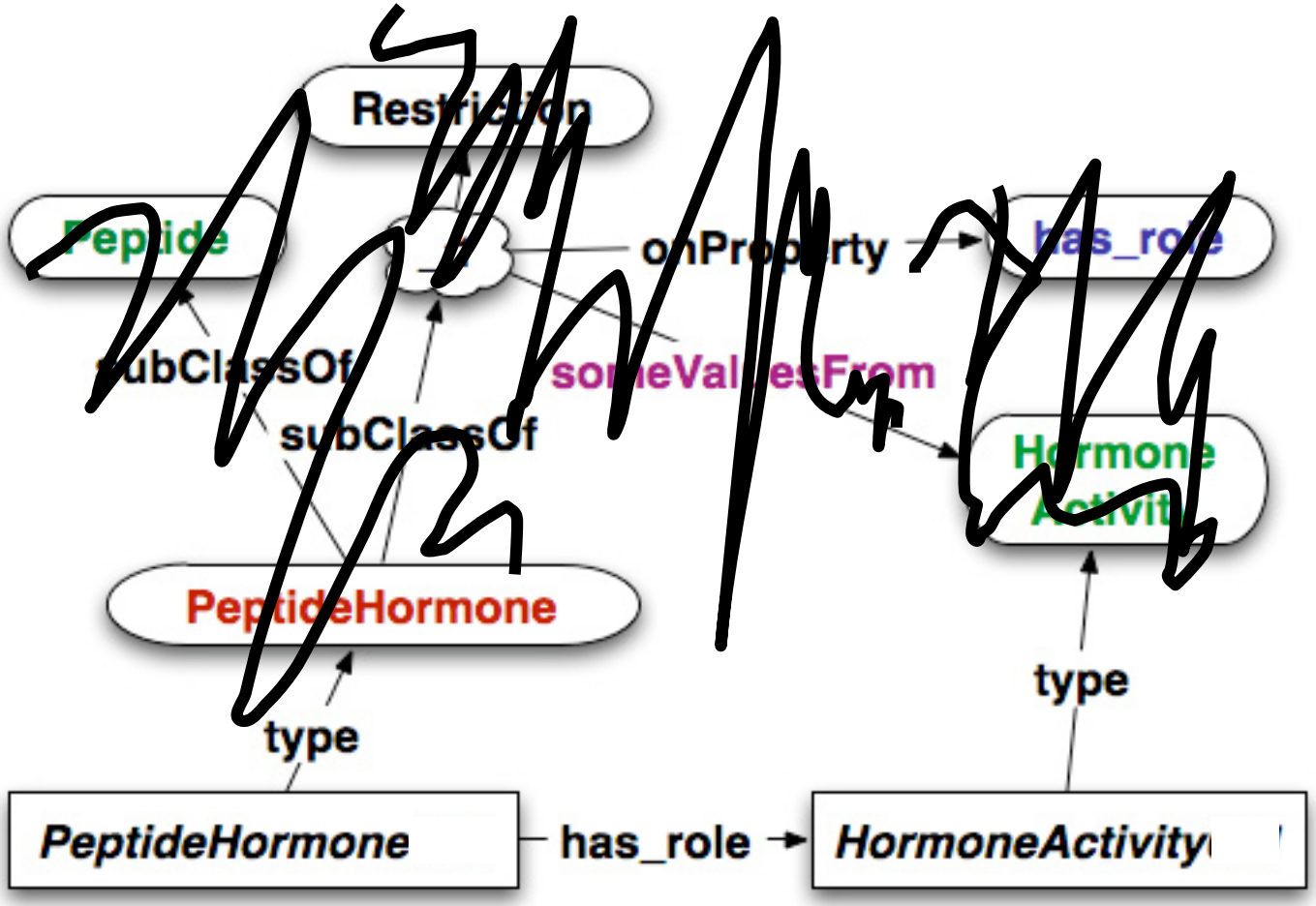
6 :_1 owl:onProperty ro:hasRole.

5 :_1 owl:someValuesFrom go:GO_00179.

```
select ?moleculeClass
where {
        ?moleculeClass rdfs:subClassOf chebi:16670.
        ?moleculeClass rdfs:subClassOf ?res.
        ?res owl:onProperty ro:hasRole.
        ?res owl:someValuesFrom go:GO_00179.
      }
```

→ ?moleculeClass = chebi:25905

Ack! Too much junk! Why not just use instances?



Why not instances? Hormone Activity from GO Molecular Function

- all : all [439998]
  - GO:0003674 : molecular_function [286402]
    - GO:0005488 : binding [86988]
      - GO:0005515 : protein binding [42794]
        - GO:0005102 : receptor binding [3784]
          - **GO:0005179 : hormone activity [604]**
            - GO:0017045 : adrenocorticotropin-releasing hormone activity [2]
            - GO:0017044 : alpha-melanocyte stimulating hormone activity [0]
            - GO:0046659 : digestive hormone activity [0]
            - GO:0008613 : diuretic hormone activity [12]
            - GO:0016913 : follicle-stimulating hormone activity [4]
            - GO:0016608 : growth hormone-releasing hormone activity [16]
            - GO:0005183 : luteinizing hormone-releasing factor activity [22]
            - GO:0030354 : melanin-concentrating hormone activity [4]
            - GO:0016085 : myoinhibitory hormone activity [2]
            - GO:0016084 : myostimulatory hormone activity [2]
            - GO:0005184 : neuropeptide hormone activity [222]
              - **GO:0016087 : ecdysiostatic hormone activity [2]**
            - GO:0016521 : pituitary adenylate cyclase activating polypeptide activity [0]
            - GO:0008437 : thyrotropin-releasing hormone activity [6]

Possible answers to the “Ugly OWL RDF” problems

- Macros
- Rules (as transformation mechanism)
- ??

PARIS: Activity center analysis

Goal: Use prior knowledge to extract higher quality signal from expression data.

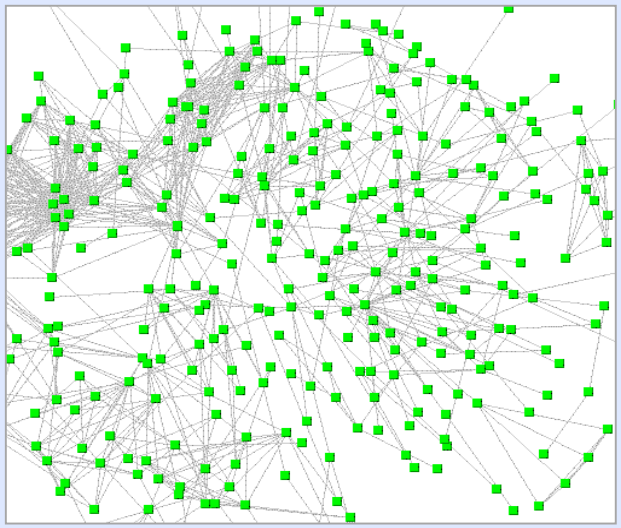
Knowledge used: Pairs of interacting proteins, as inferred from human, mouse and rat findings in KB, define a network where nodes are proteins and edges are interactions.

Strategy: Score each gene using its activity combined with activities of its neighbors; obtain P-values by testing significance; display using network layout based on distance between genes in functional network.

Method described in Pradines et al., J Biopharm. Stat., 14 (3) 2004, 701-721

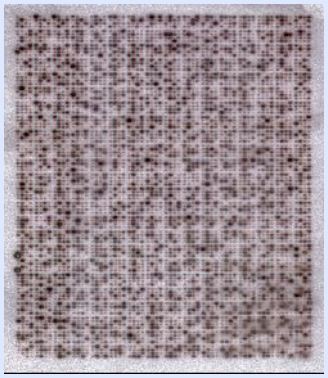
Activity center analysis

Full Interaction Network



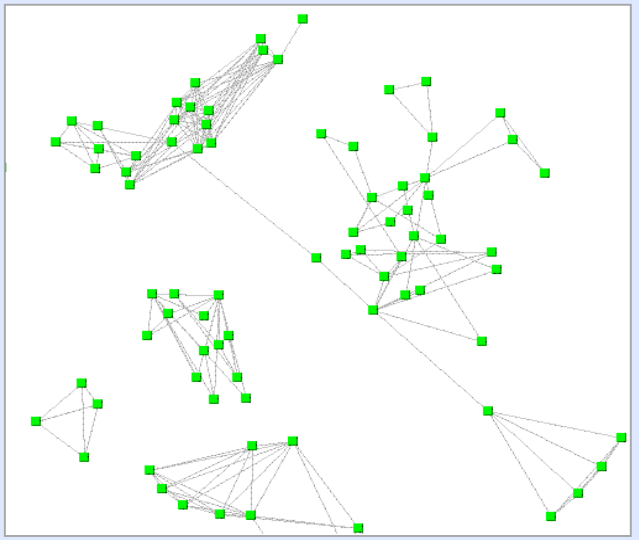
+

Data, defining *activity*



=

***Active* Sub-network**



Functional Interactions involving Gene Products

- Binds
- Phosphorylates
- Regulates
- Cleaves…

Activity

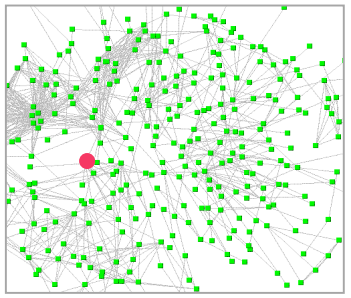
- Compound vs. Normal
- Knockout vs. Wild Type
- Responders vs. Non-responders

Hints on the Cellular Processes

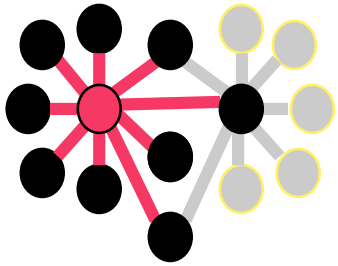
- Perturbed by a compound
- Downstream of a target
- Involved in drug resistance

Scoring activity

- **Compute activity score s_i for each gene in the network**

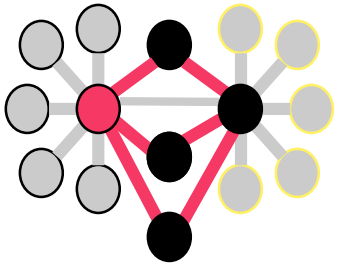


Neighborhood term a_i

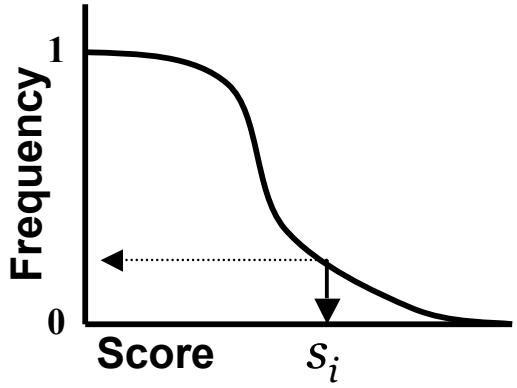


$$s_i = \frac{1}{|H|} \sum_{v_j \in H} \alpha_{ij} a_j$$

Overlap term α_{ij}

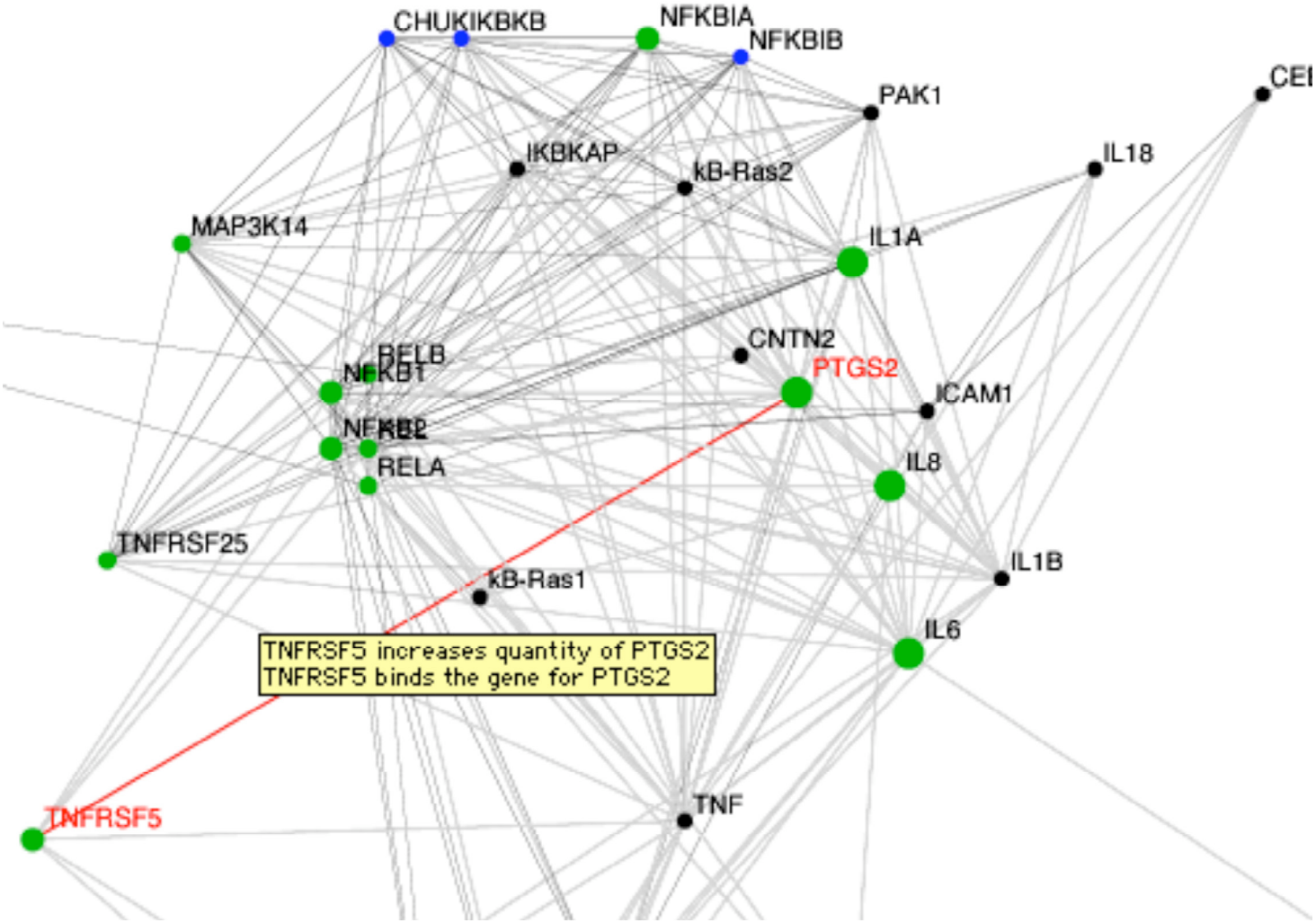


Use Monte Carlo simulation to assess significance of scores



To yield a p-value answering: how unusual is this level of activity?

Exploring an activity center in an inflammation experiment using PARIS



What we'd like to do better

Broader knowledge base - cells, anatomy, physiology, behavior, protocols, reagents

Beyond simple interaction: More precise representations of mechanism to be able to query and exploit computationally

Built in a open, scalable, scientifically credible way, to encourage sustained contribution, and to take advantage of "web effects"

How do we get there?

Interoperation is paramount, but modeling is hard: Work with the OBO Foundry

Build a skilled community

Use (open!) Semantic Web Technologies to enable web effects

Support and nurture a growing and vigorous community (SWAN, BIRN, OBI) all of whom build on the rest and enable others to build more

Work to advance key technologies and infrastructure - text mining, structured abstracts, query, reasoning.

Background Technology

So far about 350M triples in Openlink Virtuoso (~20Gb)

Commodity Hardware: 2x2core duo/2 disks/8G Ram

Biggest so far is MeSH associations to articles (200M triples)

Smaller, from 10K to 10M triples/source

A small fraction of biological knowledge!

(Don’t forget - you can still interoperate with data from relational databases)

Here's the good part!

You can play (for four more weeks at least) and download all data and install it yourself! (Thanks HP!)

Form: http://hcls1.csail.mit.edu:8890/nsparql/
Endpoint: http://hcls1.csail.mit.edu:8890/sparql/

Form: http://hcls2.csail.mit.edu:8890/nsparql/
Endpoint: http://hcls2.csail.mit.edu:8890/sparql/

More information:
http://esw.w3.org/topic/HCLS/Banff2007Demo
= http://tinyurl.com/**ywalvd**

We are actively looking for organizations to sponsor hosting this resource permanently

Acknowledgements

HCLS Demo Contributors

- John Barkley (NIST)
- Olivier Bodenreider (NLM, NIH)
- Bill Bug (Drexel University College of Medicine)
- Huajun Chen (Zhejiang University)
- Paolo Ciccarese (SWAN)
- Kei Cheung (SenseLab, Yale)
- Tim Clark (SWAN)
- Don Doherty (Brainstage Research Inc.)
- Kerstin Forsberg (AstraZeneca)
- Ray Hookaway (HP)
- Vipul Kashyap (Partners Healthcare)
- June Kinoshita (AlzForum)
- Joanne Luciano (Harvard Medical School)
- Scott Marshall (University of Amsterdam)
- Chris Mungall (NCBO)
- Eric Neumann (Teranode)
- Eric Prud'hommeaux (W3C)
- Jonathan Rees (Science Commons)
- Alan Ruttenberg (Science Commons)
- Matthias Samwald (Medical University of Vienna)

HCLS Demo Contributors

- Susie Stephens (Eli Lilly)
- Tom Stambaugh (Zeetix)
- Mike Travers
- Gwen Wong (SWAN)
- Elizabeth Wu (SWAN)

Data Providers

- Judith Blake (MGD)
- Mikail Bota (BAMS)
- David Hill (MGD)
- Oliver Hoffman (CL)
- Minna Lehvaslaiho (CL)
- Colin Knep (Alzforum)
- Maryanne Martone (CCDB)
- Susan McClatchy (MGD)
- Simon Twigger (RGD)
- Allen Brain Institute

Vendor Support

- **OpenLink** - Kingsley Idehen, Ivan Mikhailov, Orri Erling, Mitko Iliev, Patrick van Kleef
- **HP** - Ray Hookaway, Jeannine Crockford