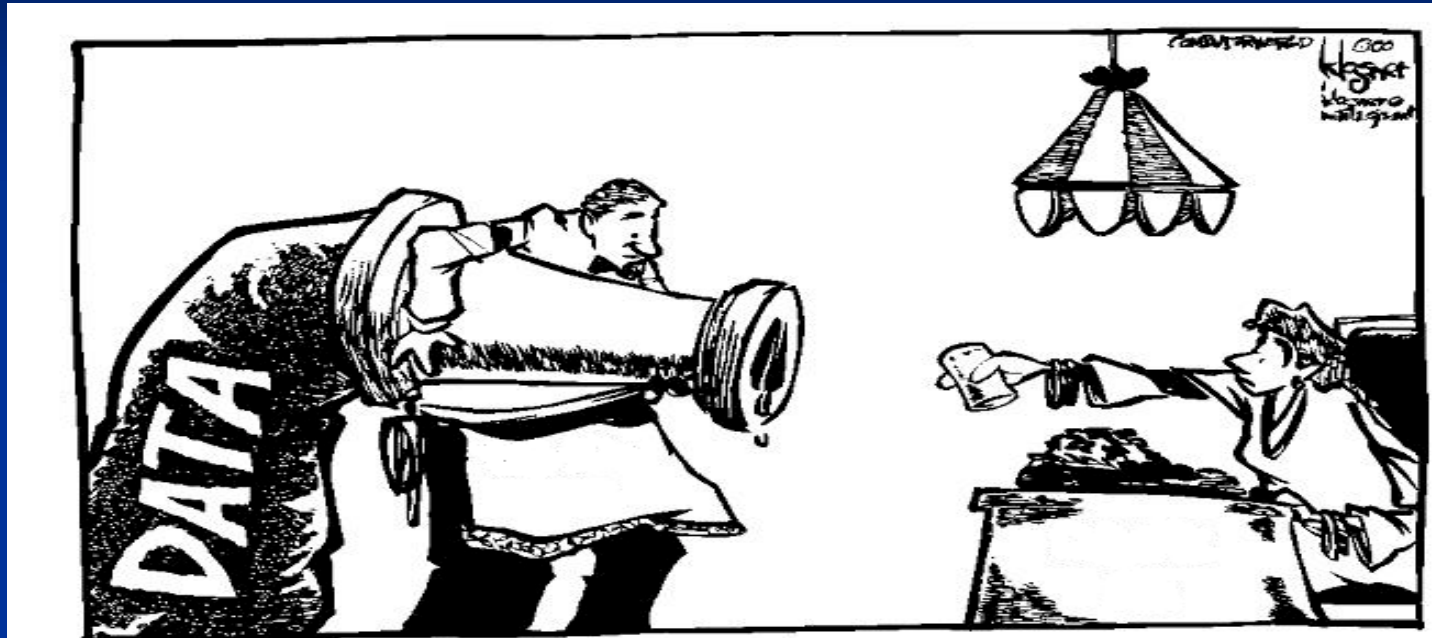


# Community Databases :: Flat files to Semantic web



Nirav Merchant  
Biotechnology Computing  
Arizona Research Laboratories  
University of Arizona  
[nirav@arl.arizona.edu](mailto:nirav@arl.arizona.edu)  
520-621-8379

# FLYBRAIN :: CA 1995 (NSF Funded)



## FLYBRAIN

An Online Atlas and Database  
of the *Drosophila* Nervous System

[Help](#)

[SiteNavigator](#)

[Index](#)

[Where Am I?](#)

[Search](#)

[Page Status](#)

[Contact Us](#)

### Contents

[Atlas of the \*Drosophila\* Brain](#)

[Genetic Dissection of the Brain](#)

[Developmental Studies](#)

[The Dissectable Brain](#)

[3D Models of the Brain](#)

[Summary Diagrams of the Brain](#)

[What's New?](#)

[Documentation and Help](#)

[User Survey](#)

[Links](#)

[\(Poster Sessions\)](#)

[SiteNavigator Help](#)

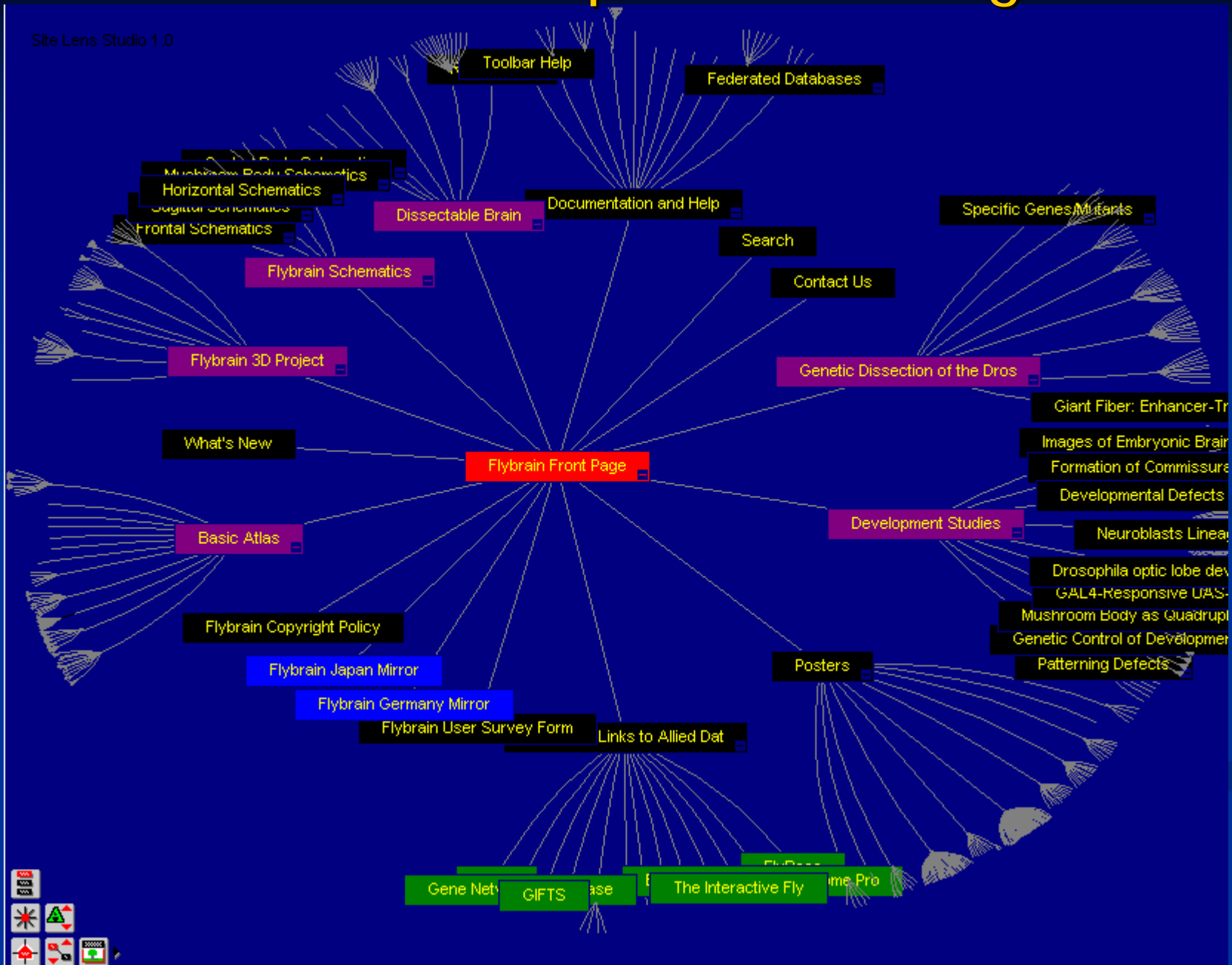
# FLYBRAIN :: Infrastructure

- Embracing open source and standards
- Operating system SGI IRIX (*1997:LINUX*)
- Web server NCSA (*1997:Apache*)
- VRML (Virtual Reality Markup Language) for 3D objects (*from 1996*)
- PERL for CGI/server side scripts
- Java for interactive browsing (*from 1996*)
- Glimpse for searching static html pages
- Meta data was stored in flat text files (tab delimited)
- 1996 mSQL (Hughes Technology, Australia) for submission tracking and work flow management

# FLYBRAIN :: Limitations

- **Data ownership** :: Pre and post publication
- **Data ownership** :: Supporting Evidence
- **Search** :: Lack of Meta information
- **Search** :: Annotations lacked controlled vocabulary
- **Search** :: Formats (Images, 3D structures)
- **Search\*** :: Spelling errors, conflicting terms
- **Data submission** :: Tedious, non standard
- **Data submission** :: Standardization on file format, size, resolution added significant work load
- Browsing many pages without a strong search engine limited use for research (worked well for instructional purposes)

# FLYBRAIN :: Graphical Browsing



# SALVIAS :: Solving ownership issues

About SALVIAS

Data

Participants

Research

Application info

Links and Sources



*Synthesis and Analysis of Local Vegetation Inventories Across Scales*

## What is SALVIAS?

### SALVIAS is:

- ◆ A web-based utility for compiling data on diverse aspects of plant organismal biology, including taxonomy, demography, phenology, and biogeography.
- ◆ Global in scope (but coverage is currently strongly biased toward the Central and South American tropics)
- ◆ A source of the following types of data ([view current SALVIAS holdings](#)):
  - ◇ local plant inventories
  - ◇ tree plots, vegetation cover plots, local species lists
  - ◇ compiled into a single, standardized database
- ◆ herbarium specimens
  - ◇ samples of individual plants from herbarium databases worldwide
  - ◇ includes complete collection information and geocoordinates, when available
- ◆ plant nomenclature
  - ◇ parsing and correction of orthographic errors
  - ◇ matching of names to world lists of plant names (currently Tropicos, Index Kewensis, Gray Card Index, and Australian Plant Names Index)
  - ◇ compilation of lists of names linked to a given name submitted (including, but not limited to, synonyms and basionyms)

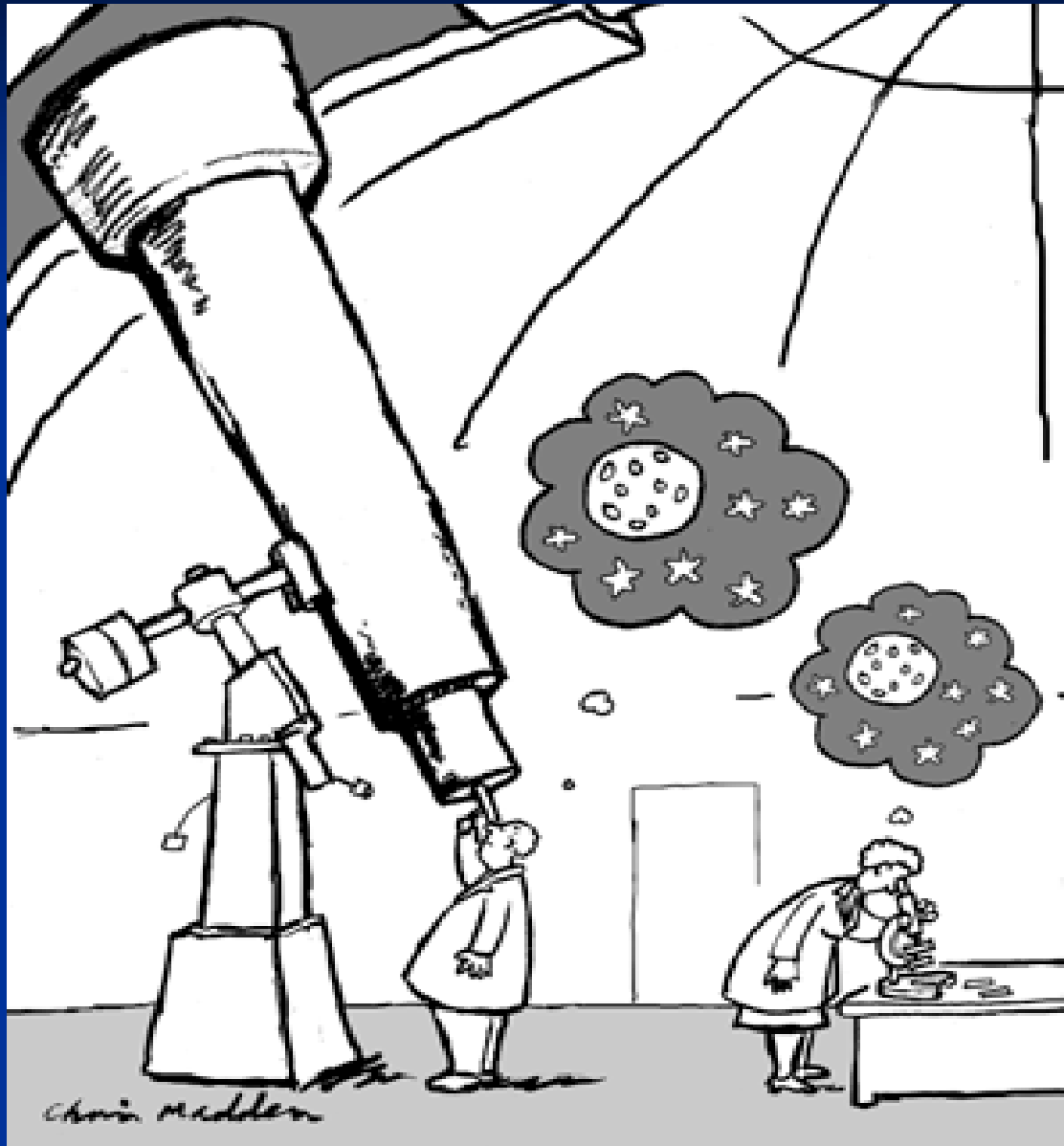
# SALVIA :: Issues

- Data submission :: Contributors wanted customized plugins for home grown systems for ease of data submission
- Nomenclature :: Varied between laboratories, institutes and countries
- Allow users to import “controlled vocabulary” into home grown solution (ended with maps)
- Needed to develop “scrubber” to clean data
- Data ownership was preserved, allowing users to search meta data, requesting author for original data if needed

# Few Distilled Message from users

- Give me a “E-lab notebook” that works
- Make me catalogue my data once and only once
- Submission to central repositories (local, global etc.) should interface with my work flow
- I hate standards, they are affecting our productivity (and keep on changing)
- Associate “my data” with “other” databases
- Very difficult to provide “raw data” for supporting evidence (too many locations and files)
- Better way to share data between collaborators\*
- Search like google (without the ads)

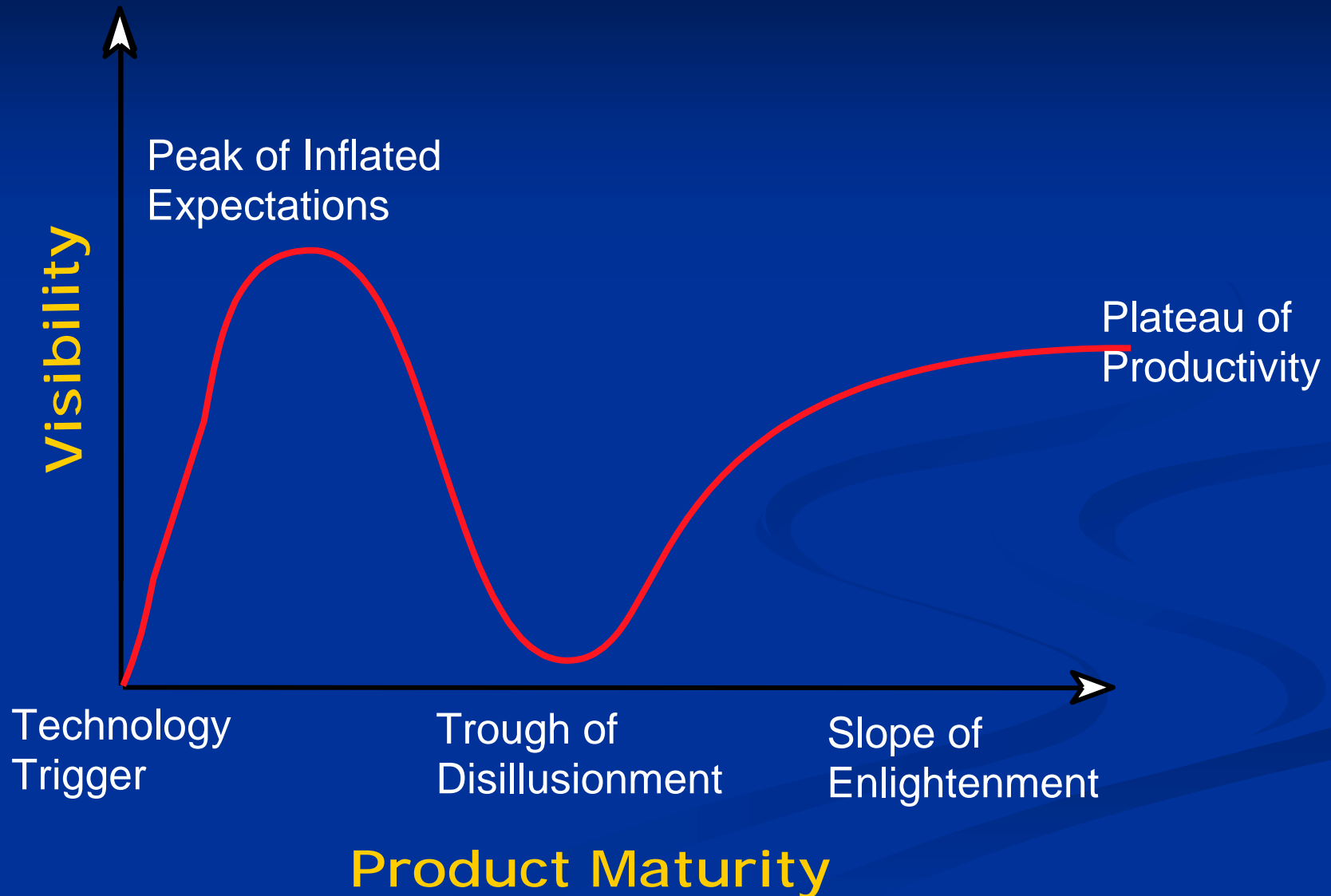
# Looking for solutions



# Semantic Web

- Provides common framework
- Allowing data to be shared and reused across application, enterprise and community boundaries
- W3C standard
- Based on RDF (Resource Description Framework)
- RDF provides a light weight ontology to facilitate exchange of knowledge
- Uses URI (Uniform Resource Identifiers, aka URL)
- Uses XML (Extensible Markup Language)
- But all of this is for machines !!!!

# Technology Hype Cycle



# Tools to build ontologies

- Protégé: <http://protege.stanford.edu>
- SWOOP: <http://www.mindswap.org> (Univ of Maryland)

## Welcome to the Protégé Project

### Protégé-2000



[What is it?](#)



[How do I get it?](#)



[How do I use it?](#)



[How do I participate?](#)



[How do I extend it?](#)

**Protégé Short Course**  
January 18th-21st, 2005  
Stanford, California

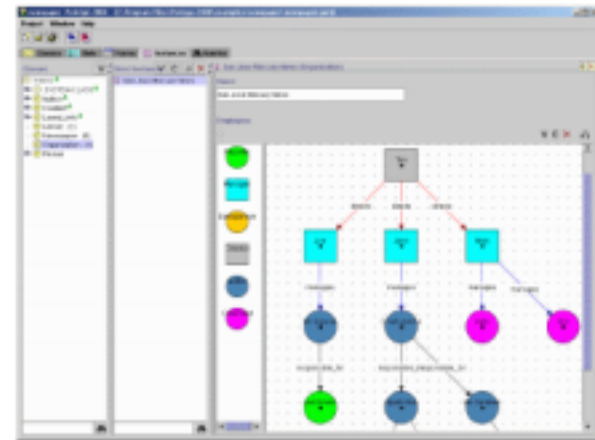


Protégé is an ontology editor and a knowledge-base editor.

Protégé is also an open-source, Java tool that provides an extensible architecture for the creation of customized knowledge-based applications.

Protégé's [OWL Plug-in](#) now provides support for editing Semantic Web ontologies.

Release 2.1.2 August 4, 2004  
Date 2.0 November 24, 2004



### Protégé Community Statistics

Registered Users	22857
users list members	9778
discussion list members	2822
discussion list messages	11421
Plug-ins	89

Updated December 9, 2004

**SUBCLASS RELATIONSHIP**

For Project: ● biopax-level1

**Asserted Hierarchy**

- owl:Thing
  - entity
    - interaction
      - control
        - catalysis**
        - modulation
      - conversion
        - biochemicalReaction
          - transportWithBiochemicalReaction
        - complexAssembly
      - transport
        - transportWithBiochemicalReaction
    - pathway
  - physicalEntity
    - complex
    - protein
    - rna
    - smallMolecule
- utilityClass
  - bioSource
  - chemicalStructure
  - dataSource
  - openControlledVocabulary

**CLASS EDITOR**

For Class: ● catalysis (instance of owl:Class)

Name:

SameAs:

DifferentFrom:

**rdfs:comment**

A control interaction in which a physical entity (a catalyst) increases the rate of a conversion interaction by lowering its activation energy. Instances of this class describe a pairing between a catalyzing entity and a catalyzed

**Annotations**

Property	Value	Lang
rdfs:comment	A control interaction in whi...	

Asserted Inferred

- Asserted Conditions**
- NECESSARY & SUFFICIENT
  - NECESSARY
  - control
  - ⊖ CONTROL-TYPE  $\exists$  "ACTIVATION"
  - ∀ CONTROLLED conversion
  - ≤ CONTROLLER ≤ 1
  - = DIRECTION = 1
  - INHERITED
  - ≤ CONTROL-TYPE ≤ 1 [from control]
  - ≤ NAME ≤ 1 [from entity]
  - ∃ PARTICIPANTS (entity  $\sqcup$  physicalEntityParticipant) [from int...]
  - ≤ SHORT-NAME ≤ 1 [from entity]

**Properties**

- COFACTOR (multiple physicalEntityPart...
- ▶ D DIRECTION (multiple Symbol)
- ▶ D CONTROL-TYPE (multiple Symbol)
- ▶ □ CONTROLLED (multiple entity, physical...
- ▶ □ CONTROLLER (multiple physicalEntityP...
- ▶ D NAME (multiple String)
- ▶ □ PARTICIPANTS (multiple entity, physica...

**Disjoints**

- modulation

# Applying semantic web to apps

- [www.mindswap.org](http://www.mindswap.org) Photostuff (overlay ontology on images)
- [www.gopubmed.org](http://www.gopubmed.org) (gene ontology + pubmed)
- <http://simile.mit.edu/> Semantic Interoperability of Metadata and Information in unLike Environments
- <http://haystack.lcs.mit.edu/> Semantic web browser !
- Semantic web for the web developer  
<http://logicerror.com/semanticWeb-webdev>

# Where is the action ?

- **W3C Workshop on Semantic Web for Life Sciences (Oct 27-28 2004)**
- **Major lesson -- data integration tools (intra-enterprise, cross-community, cross public/private boundaries) are entirely inadequate**
  - **Core Vocabularies Working Group**
  - **Investigation of identifier mechanisms and implementation strategies (LSID)**
  - **Implementers Interest Group**
- **<http://www.w3.org/2004/10/swls-workshop-report.html>**

# Its all about the data !



**LIONSHARE**

PENNSTATE

Project Information | Developers | Users | Community | Contact Information

**Project Information**

- Project News
- Project Description
- LionShare FAQ
- Documents
- Presentations
- LionShare Wiki

**Developers**

- Source Code
- Developers Wiki
- CVS Statistics
- Bug Reporting

**LionShare >> Contact Information**

Need to get a hold of the LionShare team? If you have a general question, be sure to look at the [FAQ](#). Please check our [forums](#), check our [mailing list archives](#), and blogs. If you would still like to send us a message via e-mail, our contact e-mail is [lionshare@fee.tlt.psu.edu](mailto:lionshare@fee.tlt.psu.edu).

If you are looking for a specific individual, here is a partial list of some of the people working on the LionShare project.


**LionShare Developer Directory**

Name	Position	E-Mail	Instant Messenger
Mike Halm	Project Leader	<a href="mailto:mjh@psu.edu">mjh@psu.edu</a>	<a href="#">MikeHalm</a>

- Uses P2P technology
- Allows controlled, authenticated data sharing (shibboleth project)
- Persistent data storage is possible
- Plugins for searching meta data (ECL::pool,pond)

<http://cabio.nci.nih.gov/>

**cancer.gov**

**NATIONAL CANCER INSTITUTE**  **Center for Bioinformatics**


Home Organization Initiatives Infrastructure Support Download

NCICB > Infrastructure > Enterprise Vocabulary Services :

### News

- ◆ [NCICB User Applications Manual](#)
- ◆ [caCORE 2.1.1 released](#)
- ◆ [caBIOperl released](#)
- ◆ [caBIG Events](#)
- ◆ [2003 Jamboree Presentations](#)
- ◆ [NCICB Seminars](#)
- ◆ [Publications](#)

Quick Links



## NCI Enterprise Vocabulary Services (EVS)

**Distribution List.** Up to date information and messages concerning the EVS are sent out to a distribution list hosted by the NIH. Please subscribe to the list at <http://list.nih.gov/archives/ncievs-l.html>

**Introduction.** The NCI EVS is set of services and resources that address NCI's needs for controlled vocabulary. The EVS Project is a collaborative effort of the Center for Bioinformatics and the NCI Office of Communications. The **NCI Thesaurus**, which is a biomedical thesaurus created specifically to meet the needs of the NCI, is produced by the NCI EVS project. The NCI Thesaurus is provided under an [open content license](#). The EVS Project also produces the **NCI Metathesaurus**, which is based on [NLM's Unified Medical Language System Metathesaurus](#) supplemented with additional cancer-centric vocabulary. In addition the EVS Project provides NCI with licenses for [MedDRA](#), [SNOMED](#), [ICD-O-3](#), and other proprietary vocabularies.

**Documentation.** As part of the caCORE 2.0 release, we are publishing a Technical Guide and a User's Guide containing material that covers all of [caCORE](#) including an overview of the EVS and information on how to download the September 2003 version of the NCI Thesaurus in [Ontology Web Language \(OWL\)](#), [Extensible Markup Language \(XML\)](#) or flat file format.

- Provides complete framework
- Controlled vocabulary, ontology
- Webservice (SOAP, XML-RPC)
- Java, PERL API for data access

# Using NCBI Webservices (MESH)

File Edit View Go Bookmarks Tools Help

http://lifescience.arizona.edu/search.php?string=moth&submit=go

Customize Links Free Hotmail RealPlayer Toshiba Access Windows Media Windows Enabling High Perform... Slashdot | Fighting S... MarginalHacks album...

## Graduate Studies in the LIFE SCIENCES

THE UNIVERSITY OF ARIZONA

search the database  go

### Search Results :: moth

Search using related words

allatotropin go

- allatotropin
- antheraea
- bombyx
- bombyx mori
- cecropia plant
- eclosion hormone
- giant silkmoths
- giant silkworms
- manduca
- manduca sexta
- moths
- os-d protein, drosophila
- proteins, silk gum
- sericin
- sericin 1
- sericin-1
- sericins
- silk gum proteins
- silkmoth
- silkmoths

0 Departments 7 Faculty 223 Publications

the relevancy of the search term in the faculty member's name and research description.

cellular mechanisms of physical and chemical carcinogenesis; cellular oncogene activation and differential or progression.

ons: Cell Biology & Anatomy, Molecular & Cellular Biology, Pharmacology & Toxicology, Biochemistry & Molecular Biophysics

Cell Biology & Anatomy, Cancer Biology, Molecular and Cellular Biology, Pharmacology &

Toxicology

Home

Research Areas

Programs

Spotlight

UA Life

Info Request

Facilities

Funding

UA Home

Today @ the UA

How do I...

- Find Information?
- Search Faculty?
- Search Research?

# ALICE :: Arizona Lifesciences Information Cataloguing Environment

