User-Centered Biological Information Location by Combining User Profiles and Domain Knowledge

Xindong Wu

Department of Computer Science

University of Vermont

Burlington, Vermont

Outline

1. User-Centered Agents for Information Location
2. An Intelligent Digital Library System on the Web for Computational Biologist

So much data, so little time

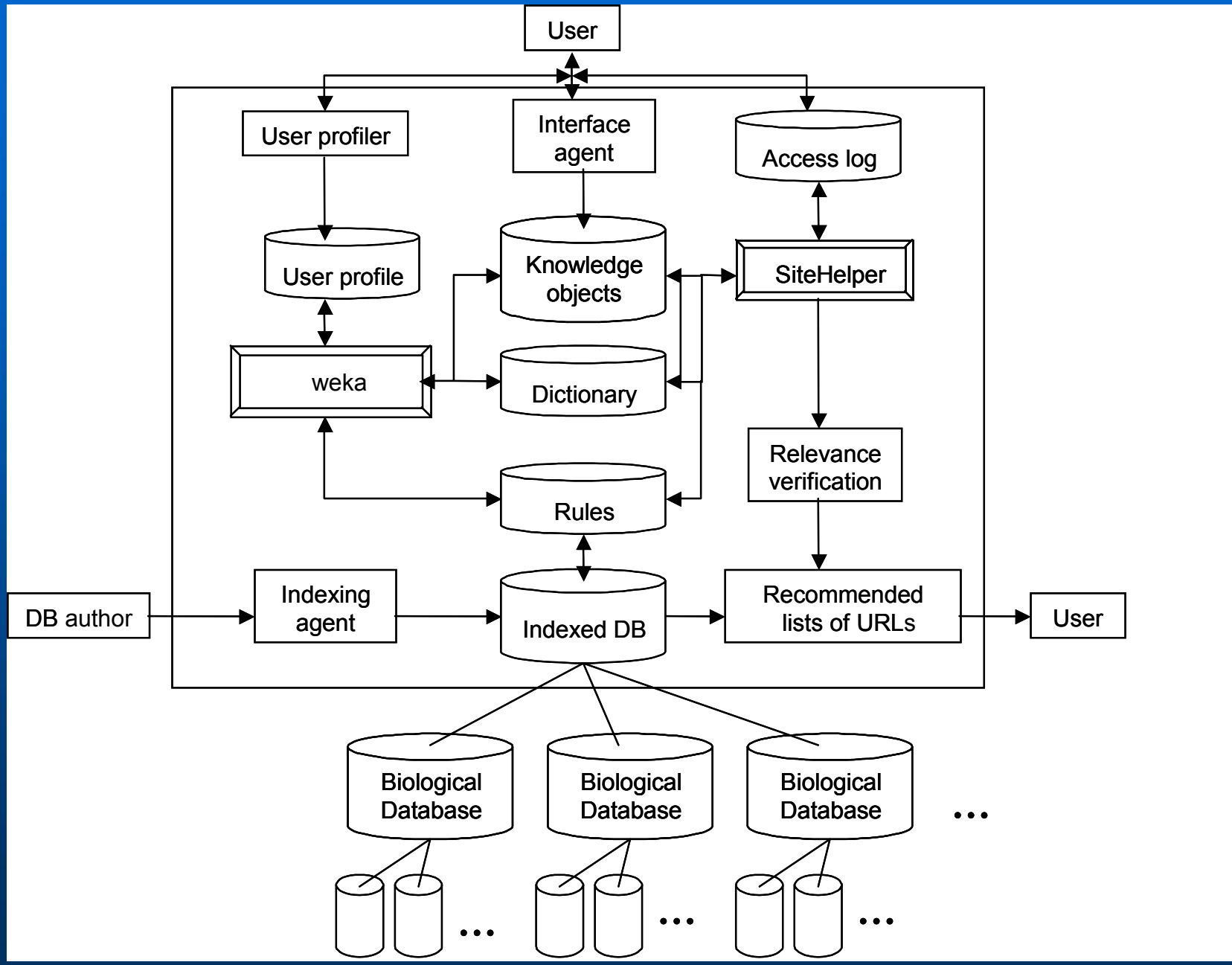
- 600,000 bio-medical research papers published per year.
- Biological papers are data rich.
- Enormous public data repositories allow users to search through this data.
- Most of this data is available via the Internet.

How can we control this data

- Large, multi-database sites such as NCBI and EMBL.
- Good search engines.
- But this requires detailed knowledge of what users are looking for.
- Takes a lot of valuable time.

Intelligent Digital Libraries

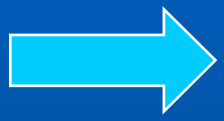
- Use the model of conventional library
- Data resides on the Internet with links to our digital library
- User profiles and intelligent agents to help guide search.


User
User profiler
Interface agent
Access log
User profile
Knowledge objects
SiteHelper
weka
Dictionary
Relevance verification
Rules
DB author
Indexing agent
Indexed DB
Recommended lists of URLs
User
Biological Database
Biological Database
Biological Database

User Profiles

- Access log collects information on what sites were visited or documents were viewed
- The user profile is updated to show this history at the end of each session
- The user profiler uses the weka data mining program to develop “rules” based on this history.

Generation of User Rules

- XML user profile  data file
- Data files are used by the (weka) data mining software
- Classification (J48 etc.) generates decision trees which we can convert into if-then type statements
- These rules are then used to recommend items in the library to the user.

Outline

1. User-Centered Agents for Information Location
2. An Intelligent Digital Library System on the Web for Computational Biologist

Semantic Network as a Dictionary

- Biology is a knowledge-based discipline
- Potential problems in representation of data:
  - Biological objects rarely have a single function
  - Function often depends on a biological state
  - Several different names often exist for the same entity.
- Semantic networks can overcome these problems and are a common type of machine-readable dictionaries

Example of Semantic Network:

WordNet: http://www.cogsci.princeton.edu/~wn/

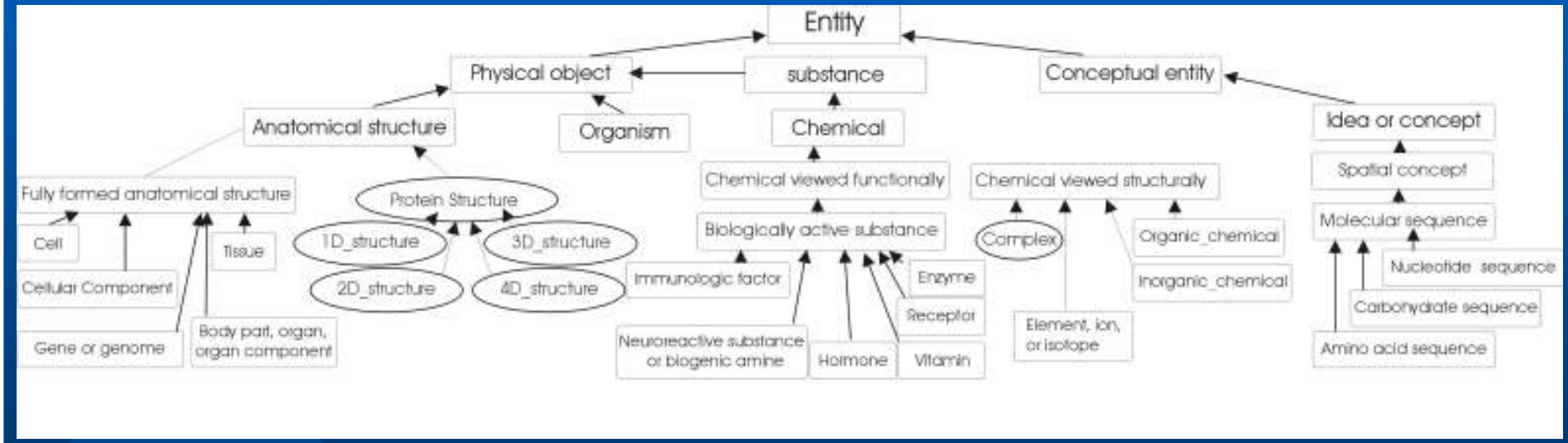
Semantic Network Structure

- Represented as a Directed Acyclic Graph (DAG).
- Nodes represent a general categorization of a concept.
- Concept classes reside at the nodes.
- Each node possibly containing several concept classes.
- Links to other concepts represent relationships.
- These links define the semantic neighborhood of the concept.

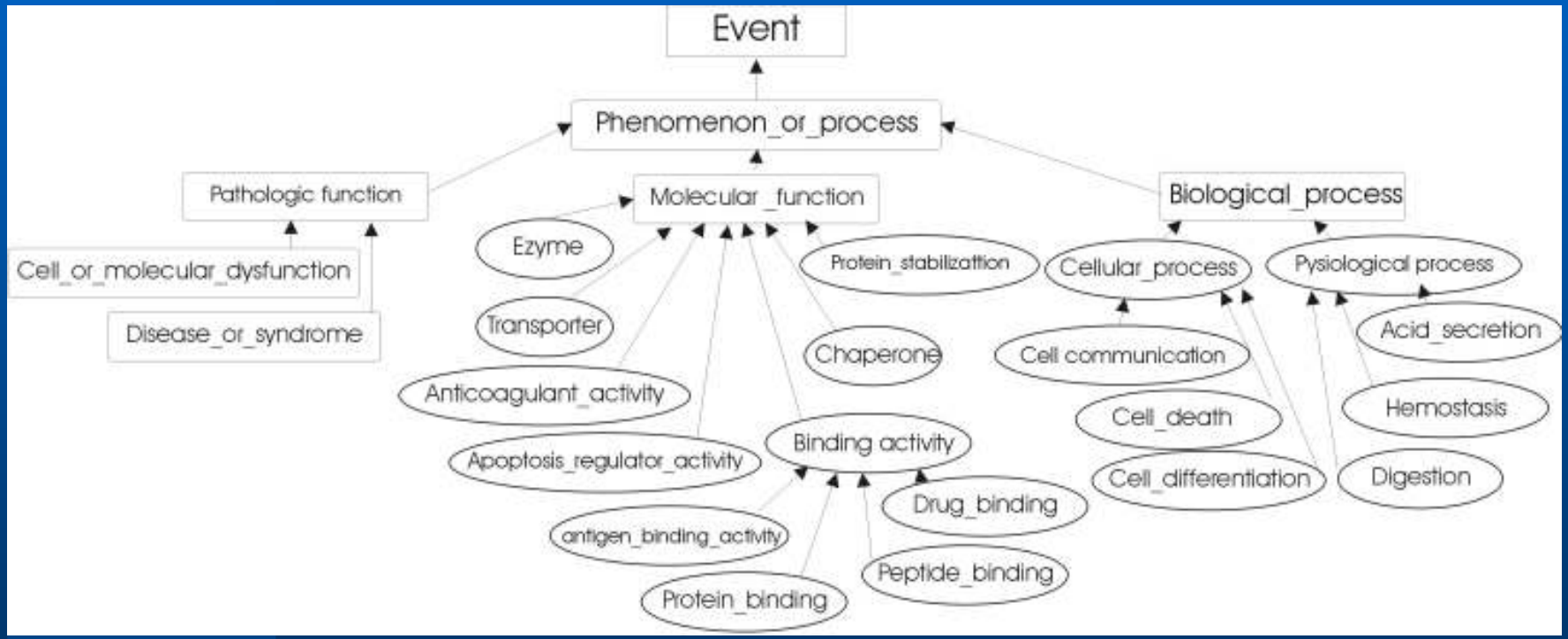
Our Semantic Network

- Based on the NLM UMLS Semantic Network
- Semantic Types are nodes that are either a biological entity, or a biological event.
  - 65 semantic types added.
  - 16 types were removed for a total of 183 nodes.
- Relationships links are either hierarchical (is-a) relationships or Associate-with relationships that link concepts together.
  - 15 new relationships for a total of 69.
- Dictionary terms reside in the concept classes at each node.

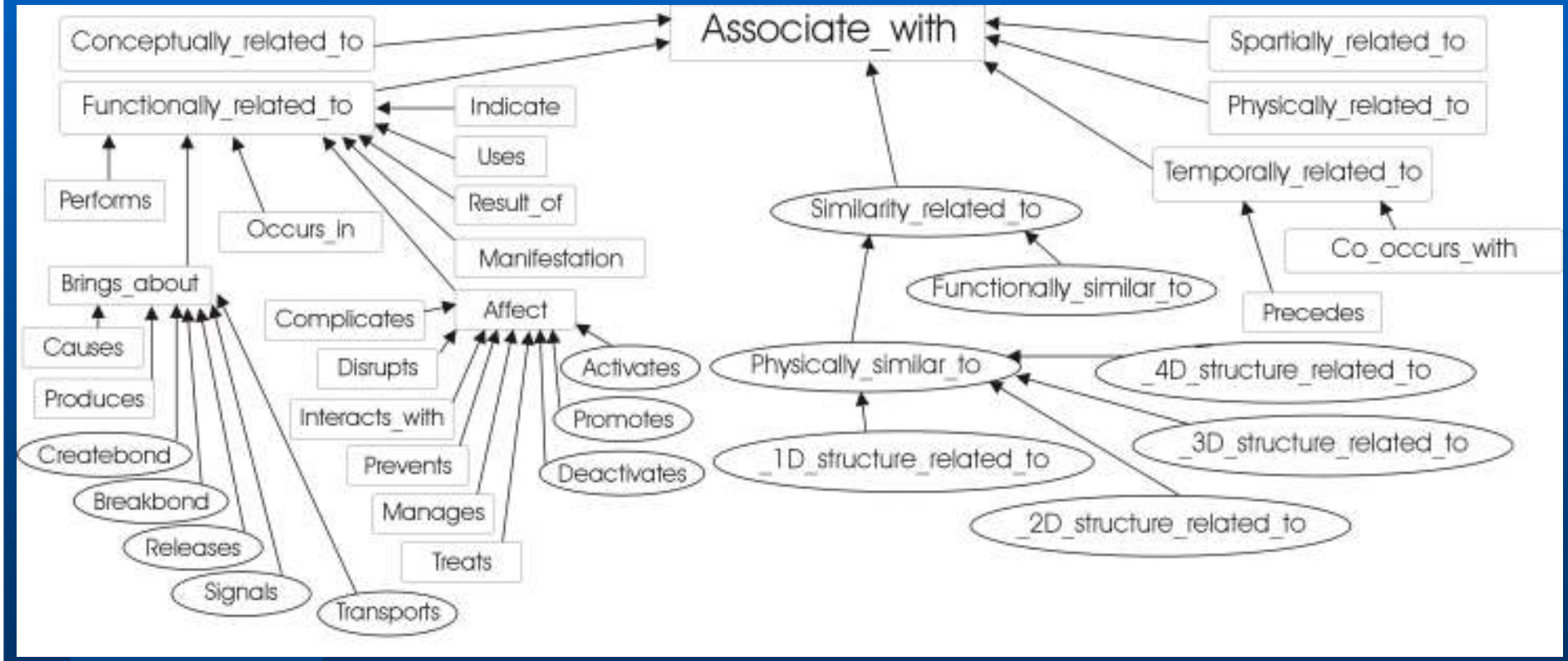
Entity Type


Entity
Physical object
substance
Conceptual entity
Anatomical structure
Organism
Chemical
Idea or concept
Fully formed anatomical structure
Protein Structure
Chemical viewed functionally
Chemical viewed structurally
Spatial concept
Cell
1D_structure
3D_structure
Biologically active substance
Complex
Organic_chemical
Molecular sequence
Tissue
Cellular Component
2D_structure
4D_structure
Immunologic factor
Enzyme
Inorganic_chemical
Nucleotide sequence
Receptor
Carbohydrate sequence
Gene or genome
Body part, organ, organ component
Neuroreactive substance or biogenic amine
Hormone
Vitamin
Element, ion, or isotope
Amino acid sequence

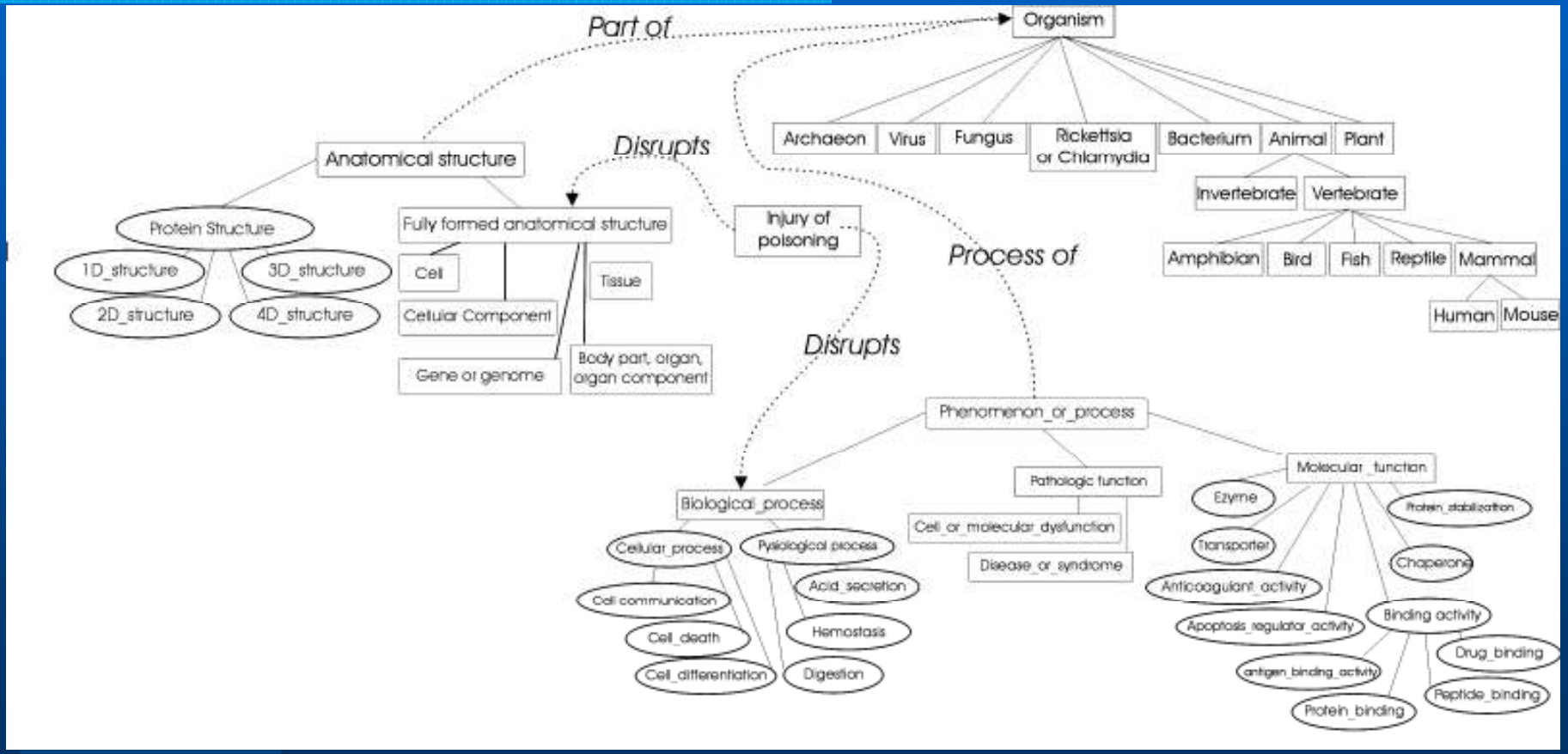
Event Type


Event
Phenomenon_or_process
Pathologic function
Molecular_function
Biological_process
Ezyme
Protein_stabilization
Cellular_process
Pysiological process
Cell_or_molecular_dysfunction
Disease_or_syndrome
Transporter
Chaperone
Cell communication
Acid_secretion
Anticoagulant_activity
Cell_death
Hemostasis
Binding activity
Apoptosis_regulator_activity
Cell_differentiation
Digestion
Drug_binding
antigen_binding_activity
Peptide_binding
Protein_binding

Relationships



Semantic Network Overview



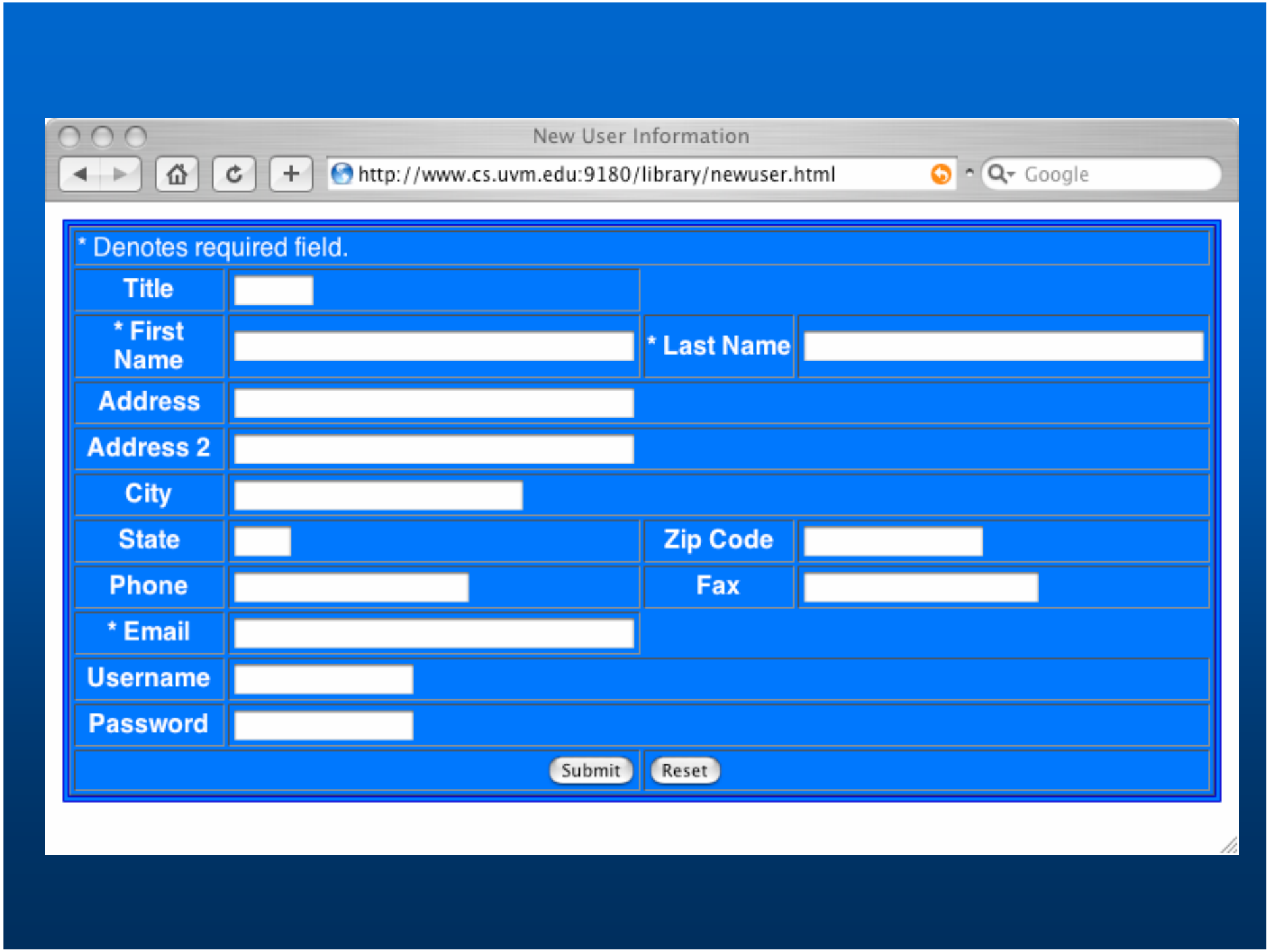
THE UNIVERSITY *of* VERMONT
Digital Library of Biological Data

User ID: jstone

Password:

Log In

New Users: Please Register , or you can use system anonymously.


New User Information
http://www.cs.uvm.edu:9180/library/newuser.html
Google
* Denotes required field.
Title
* First Name
* Last Name
Address
Address 2
City
State
Zip Code
Phone
Fax
* Email
Username
Password
Submit
Reset

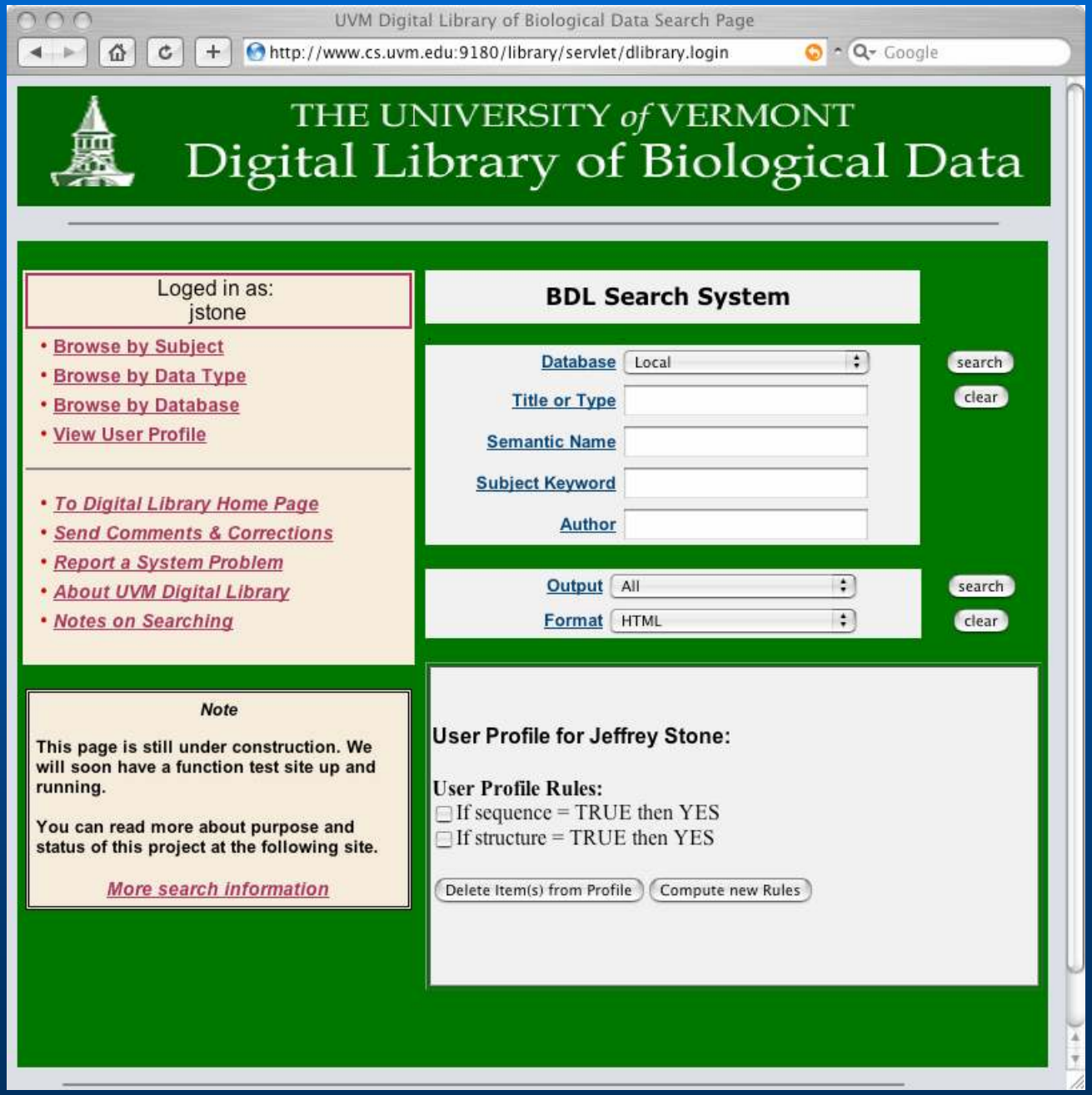
THE UNIVERSITY *of* VERMONT
Digital Library of Biological Data

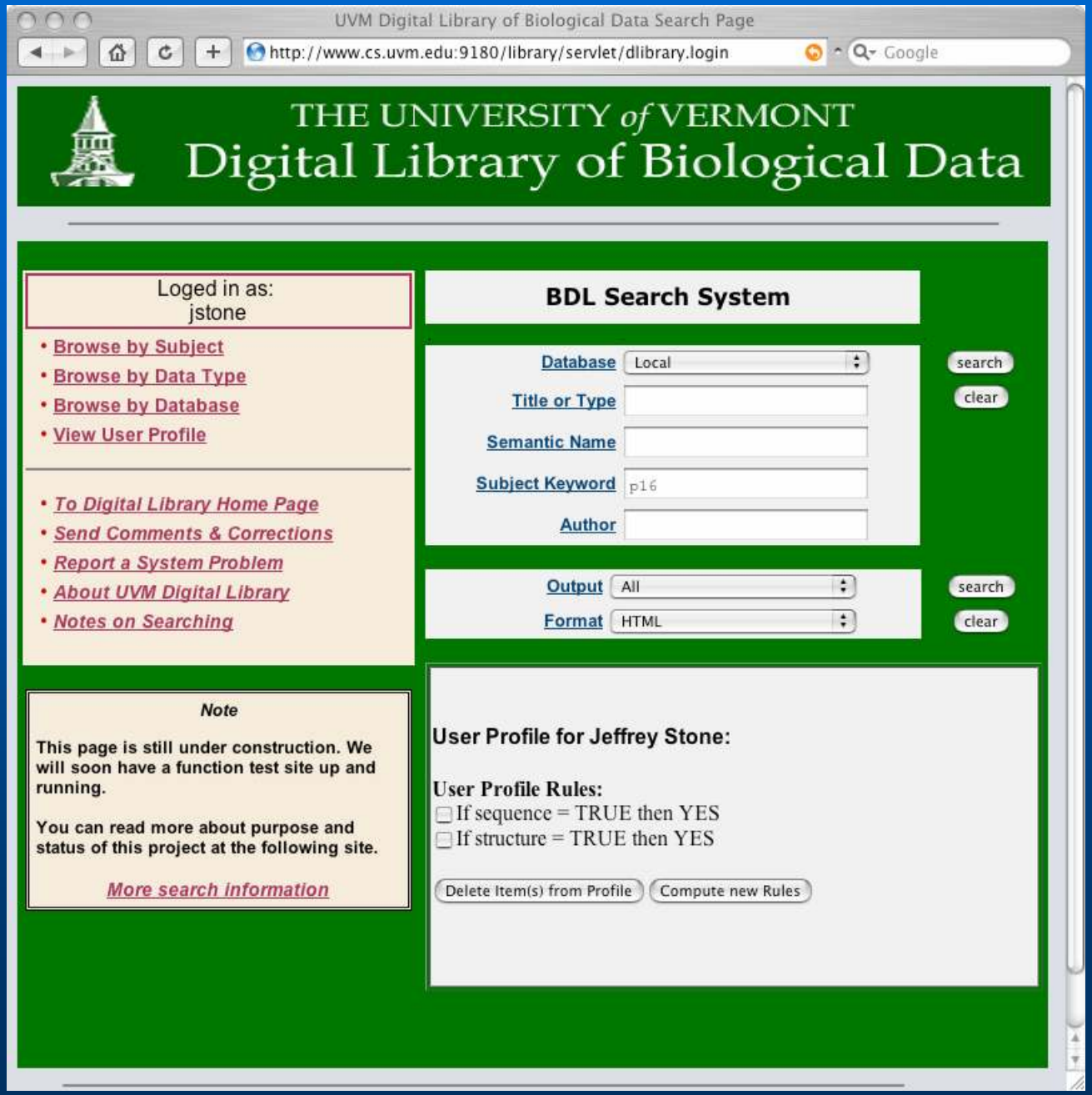
User ID: jstone

Password:

Log In

New Users: Please Register , or you can use system anonymously.


UVM Digital Library of Biological Data Search Page
http://www.cs.uvm.edu:9180/library/servlet/dlibrary.login
Google
THE UNIVERSITY of VERMONT
Digital Library of Biological Data
Loged in as:
jstone
• Browse by Subject
• Browse by Data Type
• Browse by Database
• View User Profile
• To Digital Library Home Page
• Send Comments & Corrections
• Report a System Problem
• About UVM Digital Library
• Notes on Searching
Note
This page is still under construction. We will soon have a function test site up and running.
You can read more about purpose and status of this project at the following site.
More search information
BDL Search System
Database
Local
Title or Type
Semantic Name
Subject Keyword
Author
search
clear
Output
All
Format
HTML
search
clear
User Profile for Jeffrey Stone:
User Profile Rules:
If sequence = TRUE then YES
If structure = TRUE then YES
Delete Item(s) from Profile
Compute new Rules


UVM Digital Library of Biological Data Search Page
http://www.cs.uvm.edu:9180/library/servlet/dlibrary.login
Google
THE UNIVERSITY of VERMONT
Digital Library of Biological Data
Loged in as:
jstone
• Browse by Subject
• Browse by Data Type
• Browse by Database
• View User Profile
• To Digital Library Home Page
• Send Comments & Corrections
• Report a System Problem
• About UVM Digital Library
• Notes on Searching
Note
This page is still under construction. We will soon have a function test site up and running.
You can read more about purpose and status of this project at the following site.
More search information
BDL Search System
Database
Local
Title or Type
Semantic Name
Subject Keyword
p16
Author
search
clear
Output
All
Format
HTML
search
clear
User Profile for Jeffrey Stone:
User Profile Rules:
If sequence = TRUE then YES
If structure = TRUE then YES
Delete Item(s) from Profile
Compute new Rules

Your Search Results

Click checkbox next to items you would like to add to your profiles

☐ **14 Detailed computational study of p53 and p16: using**
item type: database

☐ **12 UVM BioDesktop - CDKN2a Database Project**
item type: database

☐ **15 Detailed computational study of p53 and p16: using**
item type: database

☐ **16 p16INK4 mutations and altered expression in huma**
item type: database

Add To Profile

Items you have previously selected

☐ 26 ExPASy Proteomics Server
item type: database

☐ 15 Detailed computational study of p53 and p16: using
item type: database

☐ 12 UVM BioDesktop - CDKN2a Database Project
item type: database

☐ 14 Detailed computational study of p53 and p16: using
item type: database

☐ 22 Fadd Death Effector Domain, F25Y Mutant, NMR Minim
item type: web page

☐ 23 Fas engagement induces neurite growth through ERK
item type: web page

Delete from Profile

Additional Items matching your rule set

Rule # 1: If sequence = TRUE then YES

14 Detailed computational study of p53 and p16: using
item type: database

12 UVM BioDesktop - CDKN2a Database Project
item type: database

11 UVM BioDesktop
item type: database

15 Detailed computational study of p53 and p16: using
item type: database

17 NCBI
item type: database

26 ExPASy Proteomics Server
item type: database

Rule # 2: If structure = TRUE then YES

12 UVM BioDesktop - CDKN2a Database Project
item type: database

17 NCBI
item type: database

23 Fas engagement induces neurite growth through ERK
item type: web page

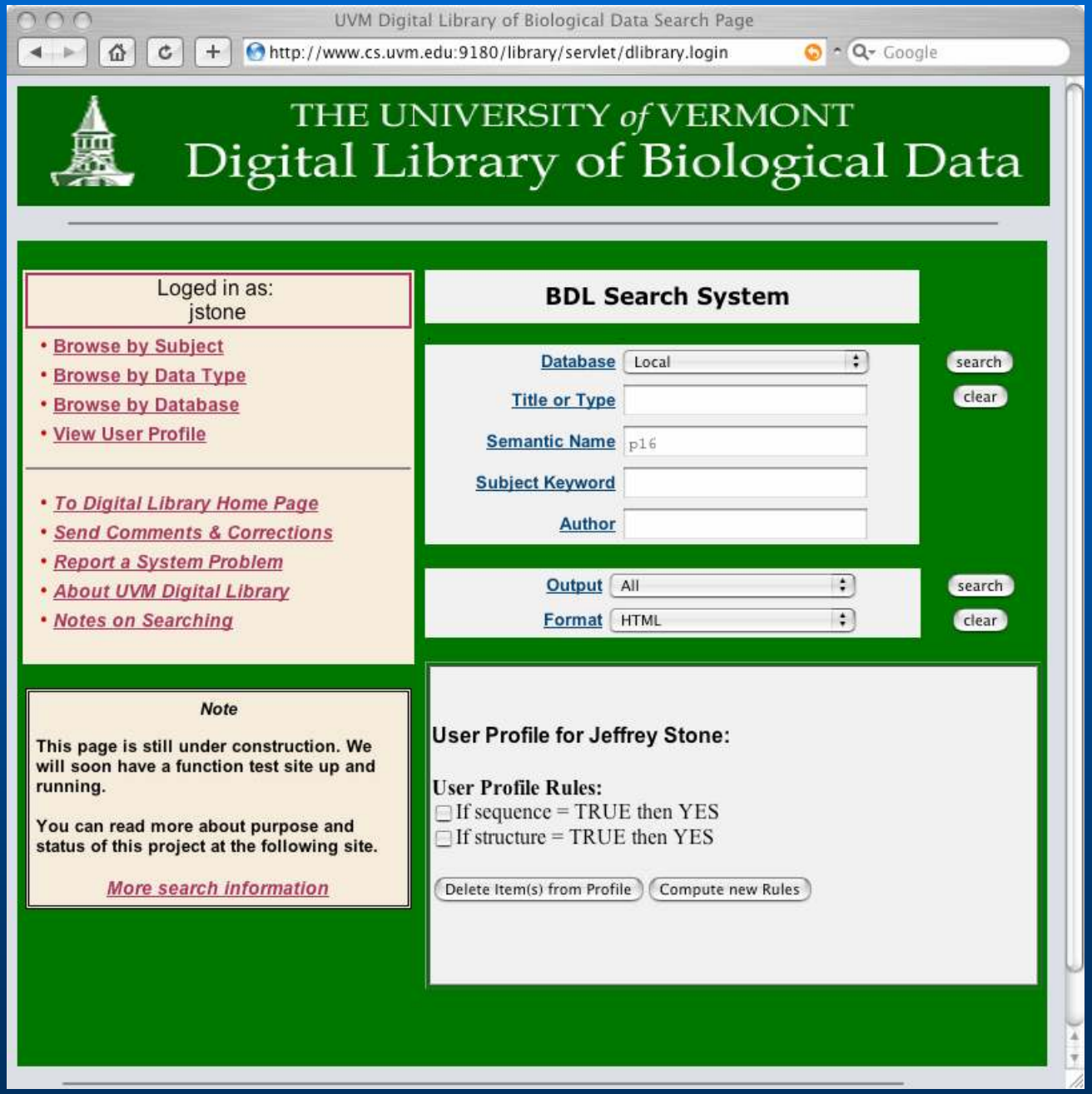
21 Protein Data Bank (PDB)
item type: database

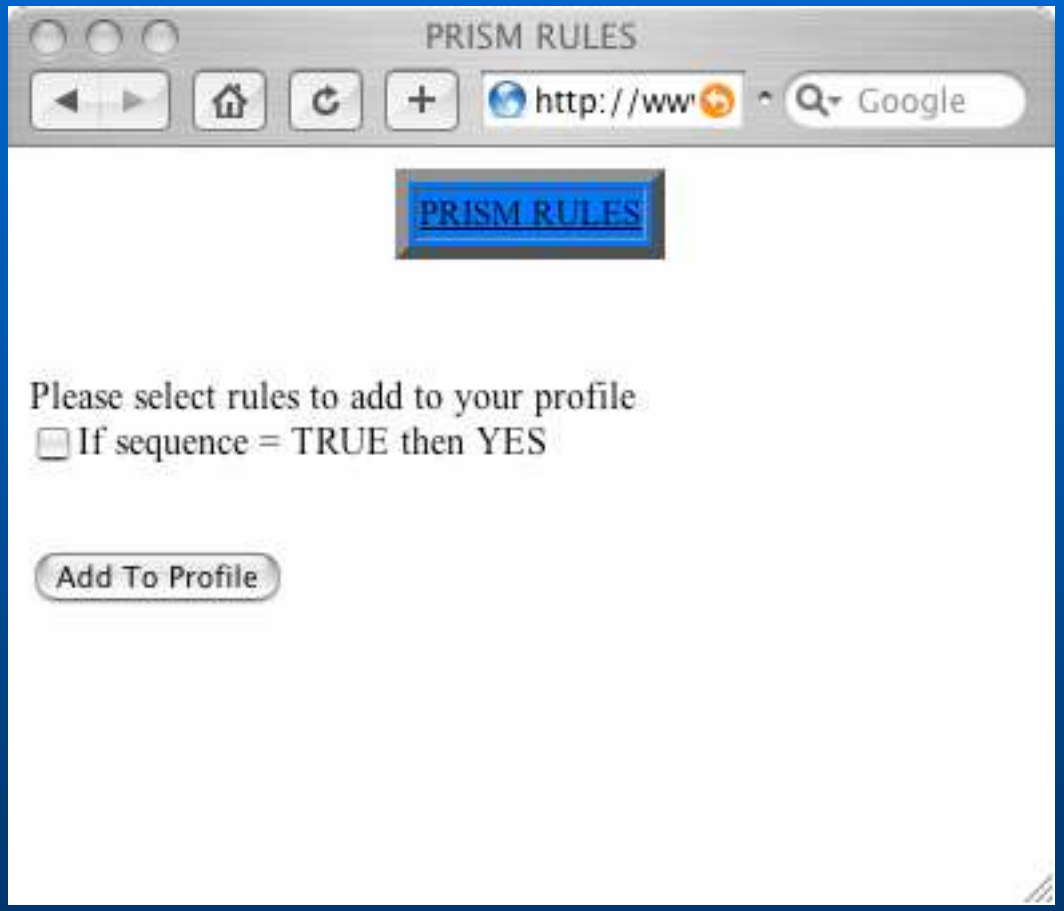
22 Fadd Death Effector Domain, F25Y Mutant, NMR Minim
item type: web page

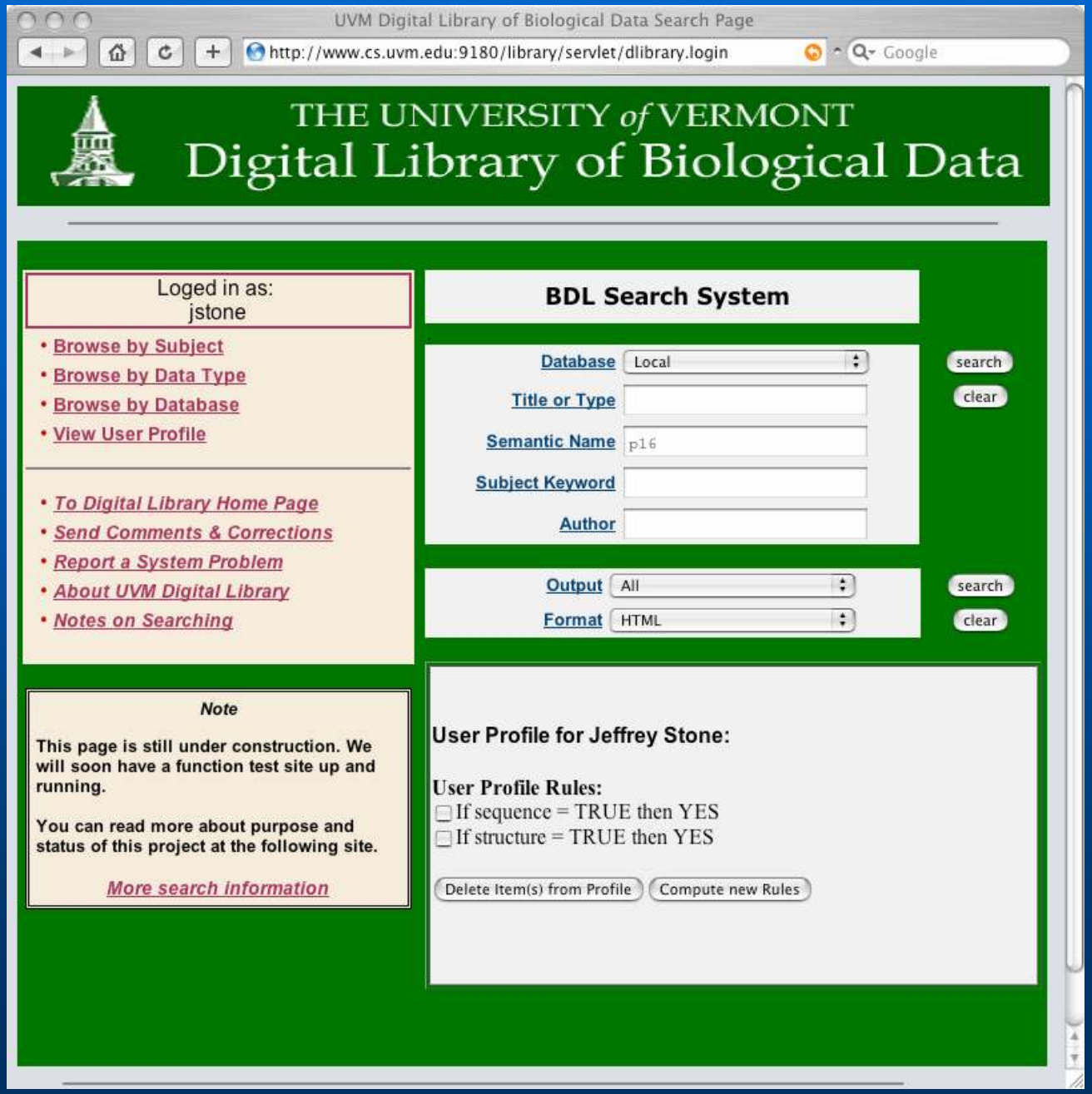
26 ExPASy Proteomics Server
item type: database

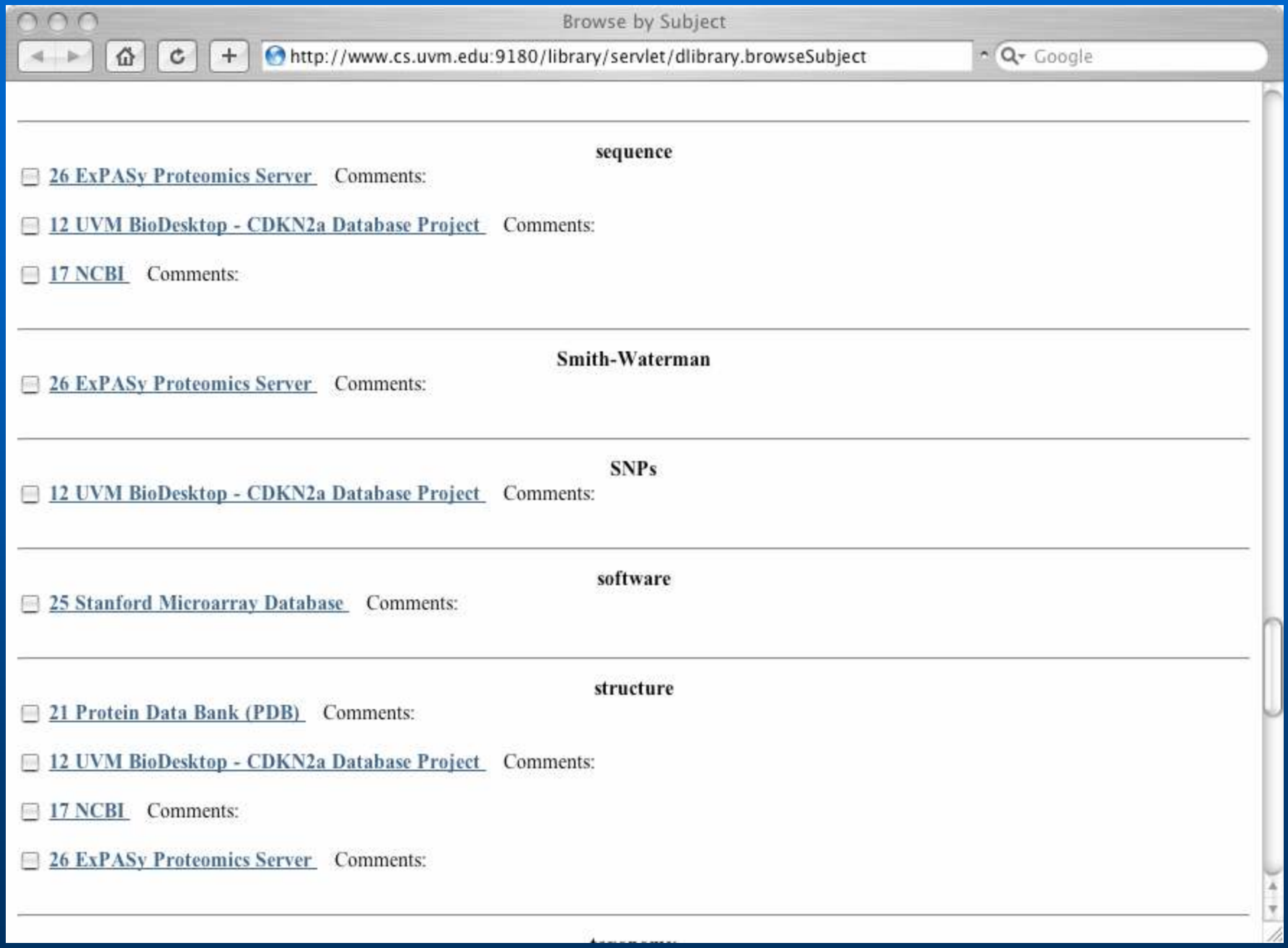
```
@relation docrules

@attribute uid {14,13,12,11,15,16,17,23,21,22,24,25,26}
@attribute sequences {TRUE, FALSE}
@attribute mutations {TRUE, FALSE}
@attribute phylogenetics {TRUE, FALSE}
@attribute web_application {TRUE, FALSE}
@attribute p16 {TRUE, FALSE}
@attribute CDKN2a {TRUE, FALSE}
@attribute p14/ARF {TRUE, FALSE}
@attribute sequence {TRUE, FALSE}
@attribute structure {TRUE, FALSE}
@attribute phylogeny {TRUE, FALSE}
@attribute SNPs {TRUE, FALSE}
@attribute cancer {TRUE, FALSE}
@attribute tumor_suppressor_gene {TRUE, FALSE}
@attribute CDKN2A {TRUE, FALSE}
@attribute protein {TRUE, FALSE}
@attribute p14 {TRUE, FALSE}
@attribute p14ARF {TRUE, FALSE}
@attribute CMM2 {TRUE, FALSE}
@attribute chromosome_9 {TRUE, FALSE}
@attribute INK4 {TRUE, FALSE}
@attribute database {TRUE, FALSE}
@attribute publications {TRUE, FALSE}
@attribute taxonomy {TRUE, FALSE}
```

```
@data
14,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,
FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FAL
SE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE, NO
13,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,
FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FAL
SE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE, NO
12,FALSE,FALSE,FALSE,FALSE,TRUE,TRUE,TRUE,TRUE,TRUE,TRUE,TRUE,TRUE,TRUE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FAL
SE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,
FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE, YES
11,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,
FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FAL
SE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE,FALSE, NO
```


UVM Digital Library of Biological Data Search Page
http://www.cs.uvm.edu:9180/library/servlet/dlibrary.login
Google
THE UNIVERSITY of VERMONT
Digital Library of Biological Data
Loged in as:
jstone
• Browse by Subject
• Browse by Data Type
• Browse by Database
• View User Profile
• To Digital Library Home Page
• Send Comments & Corrections
• Report a System Problem
• About UVM Digital Library
• Notes on Searching
Note
This page is still under construction. We will soon have a function test site up and running.
You can read more about purpose and status of this project at the following site.
More search information
BDL Search System
Database
Local
Title or Type
Semantic Name
p16
Subject Keyword
Author
search
clear
Output
All
Format
HTML
search
clear
User Profile for Jeffrey Stone:
User Profile Rules:
If sequence = TRUE then YES
If structure = TRUE then YES
Delete Item(s) from Profile
Compute new Rules


PRISM RULES
http://ww
Google
PRISM RULES
Please select rules to add to your profile
If sequence = TRUE then YES
Add To Profile


UVM Digital Library of Biological Data Search Page
http://www.cs.uvm.edu:9180/library/servlet/dlibrary.login
Google
THE UNIVERSITY of VERMONT
Digital Library of Biological Data
Loged in as:
jstone
• Browse by Subject
• Browse by Data Type
• Browse by Database
• View User Profile
• To Digital Library Home Page
• Send Comments & Corrections
• Report a System Problem
• About UVM Digital Library
• Notes on Searching
Note
This page is still under construction. We will soon have a function test site up and running.
You can read more about purpose and status of this project at the following site.
More search information
BDL Search System
Database
Local
Title or Type
Semantic Name
p16
Subject Keyword
Author
search
clear
Output
All
Format
HTML
search
clear
User Profile for Jeffrey Stone:
User Profile Rules:
If sequence = TRUE then YES
If structure = TRUE then YES
Delete Item(s) from Profile
Compute new Rules


Browse by Subject
http://www.cs.uvm.edu:9180/library/servlet/dlibrary.browseSubject
Google
sequence
26 ExPASy Proteomics Server Comments:
12 UVM BioDesktop - CDKN2a Database Project Comments:
17 NCBI Comments:
Smith-Waterman
26 ExPASy Proteomics Server Comments:
SNPs
12 UVM BioDesktop - CDKN2a Database Project Comments:
software
25 Stanford Microarray Database Comments:
structure
21 Protein Data Bank (PDB) Comments:
12 UVM BioDesktop - CDKN2a Database Project Comments:
17 NCBI Comments:
26 ExPASy Proteomics Server Comments:

Challenges

- How to structure and manage user profiles
- Provide an intelligent interface for entering and verifying items
- Facilitate semantic search:
  - Searching on relationships
  - Broadening and narrowing of keyword search
- Design an update agent
- Build composite page of content.

Thank You!

- Daniel Ngu
- Jeff Stone
- Marc Greenblatt (Medicine)

FOR MORE INFORMATION:

http://www.cs.uvm.edu/~xwu