

Science and the Semantic Web

Jim Hendler

<http://www.cs.umd.edu/~hendler>

<http://www.mindswap.org>

eScience for the rest of us!

Jim Hendler

<http://www.cs.umd.edu/~hendler>

<http://www.mindswap.org>

Cyberinfrastructure?

- NSF has a mandate to help support scientists
 - They are creating a division of "Cyberinfrastructure"
 - What will it do?

http://www.nsf.gov/news/special_reports/cyber/index.jsp

My Reply

- I would like to expand the vignette in the cyberinfrastructure web site starting with an explanation of why the astronomer was in the coffee shop that morning. You see, she had just pulled an all-nighter doing the work she needed to to get the computation ready to run. First, she had to spend a couple of hours using Google to see if she could find some programs to use for her simulation models. She then spent a couple of hours searching through the appendices of papers she found on the open-source physics archives to find the datasets that represented the particular galaxies she wanted to explore (and which, of course, cannot be searched for in Google which has no search capability against data). At that point, she had to start chatting with colleagues in Japan (who were just waking up at that time of night), because the datasets she had found were not in the format she needed for the program she needed, so she had to find a program that she could use to convert formats. Towards dawn, she had all the components she needed, unfortunately, she was writing some glue code that would create the workflow she needed to execute the whole thing. Finally, she called friends at a half dozen computer centers as they came in in the morning so she could get all the passwords and keys that would be needed so her code could run in the distributed system.

The rest of us?

- The average scientist has seen little effect from the massive funding of (mainly genomic) eScience
 - Developing an application cannot be a million dollar/project effort
 - Unless there is some replicability
 - The "fruits" of funding genomics must spin off to the rest of us
 - Tools need to support what scientists do

What Scientists do

- **Searching**, reading and thinking critically about the professional literature in their field
 - **Formulating** testable **hypotheses** consistent with the “story” or explanatory model.
 - **Finding** possible **connections** amongst disparate data, creating a plausible explanatory “story” or model which can bridge gaps or open challenges in the existing body of knowledge
 - **Designing experiments** to test their hypotheses
 - **Running the experiments**
 - **Collecting** and **analyzing** experimental data.
 - **Interpreting** data, e.g. by modifying the hypothesis, connecting it to other findings or hypotheses
 - **Organizing** personal collections of publications and related documents according to a relevant conceptual system to enable retrieval at a later date
 - **Applying for grants** to support their work (which typically involves presenting the model, hypotheses and preliminary data)
 - **Communicating** with other researchers, funding agencies, publishers, conference organizers and local institutional management
 - **Writing** scientific articles for publication, preparing conference presentations, informal talks and poster sessions.
- (Gao, Kinoshita, Wu, Lee, Miller, Clark 2005)

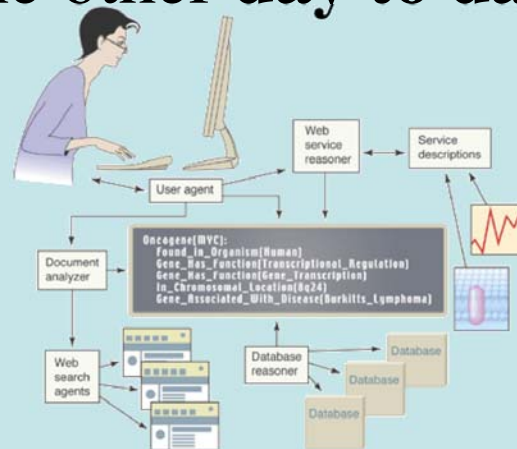
Support for eScience is support for these too!

Key to success

Tools ... must be built in a way that they **tie into the "business processes" of the working scientist** -- that is, rather than learning a whole new set of tools, the basic web tools of the scientist must include **mechanisms that make it EASIER** for the scientist ... while authoring papers, performing experiments, creating and logging data, and the other **day to day activities** of the working researcher.

Science and the Semantic Web, Hendler, 03

Cut from final article



E-science investment

- We have ignored Web lessons
 - How do we get the network effect?
- We have ignored some key issues of how scientists work
 - Models and modeling
- We have ignored some key issues of how scientists work
 - Tool embedding and the scientific process
- We have ignored some key issues of how scientists communicate
 - Jargons vs. interdisciplinary communication

Scientific impact

- Current e-science applications largely specialized to specific groups and disciplines
 - Many scientists left out
 - In e-science program “Interdisciplinary” is often used to mean CS and scientist working together
- What about chemist with physicist with cancer researcher with public policy scientist with medical doctor with ...
 - c.f. Children’s health initiative
 - c.f. Cancer risk assessments
 - c.f. Biodiversity modeling

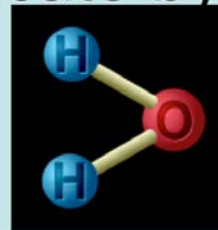
Semantic Web

- Key "next generation" evolution of the Web
 - Languages reaching maturity
 - RDF, RDFS, OWL all W3C recommendations
 - Support growing
 - Major vendors playing (Adobe, Oracle, ...)
 - Think of it as XML-tagging on steroids
 - Err, uhh, except not necessarily document based - this is a key difference!

A very old idea in new clothes

- Scientists communicate by use of models

- c.f. Physical



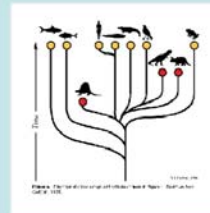
- c.f. Mathematical

Mathematical model

$$\begin{aligned} \nabla^2 \phi &= 0 \\ \eta_1 + \eta_2 \phi_1 + \eta_3 \phi_2 - \eta_4 &= 0 \\ \phi_1 + \frac{1}{2}(\phi_1^2 + \phi_2^2 + \phi_3^2) + g\eta &= 0 \\ \frac{\partial \phi}{\partial t} &= 0 \end{aligned}$$

October 21, 1998 Peng Cai

- c.f. Organizational



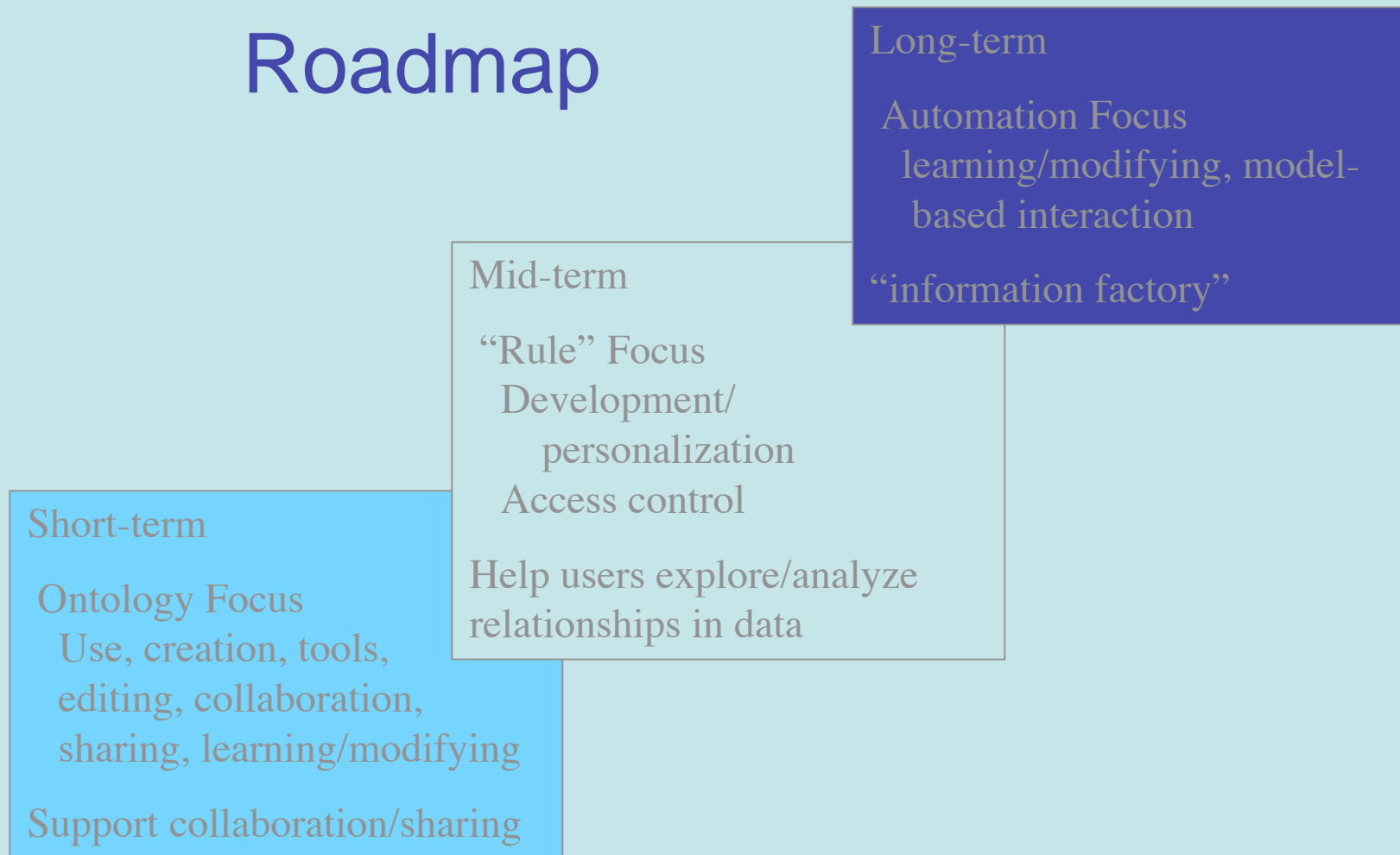
Models expose semantics

Web Modeling Languages

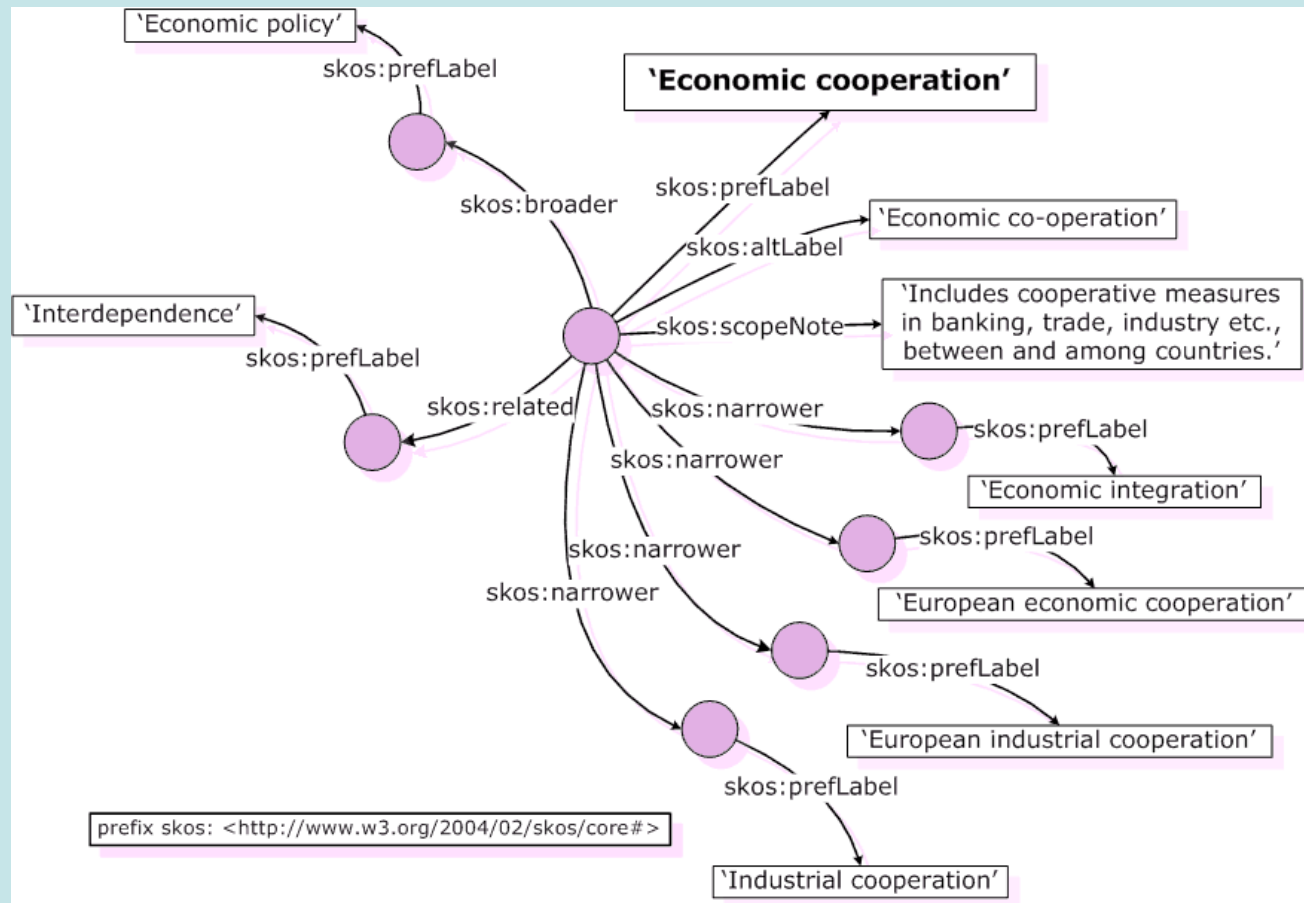
- **Resource Description Framework (RDF)**
 - ◆ Few, but important, constraint
 - ◆ A basic, extensible assertional language
- **RDF Schema (RDFS)**
 - ◆ Weak structuring of sets of terms (taxonomy-esque)
 - ◆ Class and property hierarchies
 - ◆ Domain and Range constraints
- **The Web Ontology Language, OWL**
 - ◆ Stronger structuring of sets of terms (ontologies)
 - ◆ Everything in RDFS plus
 - Complex Class constructors (unionOf, intersectionOf)
 - Additional property features (inverse, transitive)
 - Class local property type and cardinality constraints
 - And more



Sem Web Research Roadmap

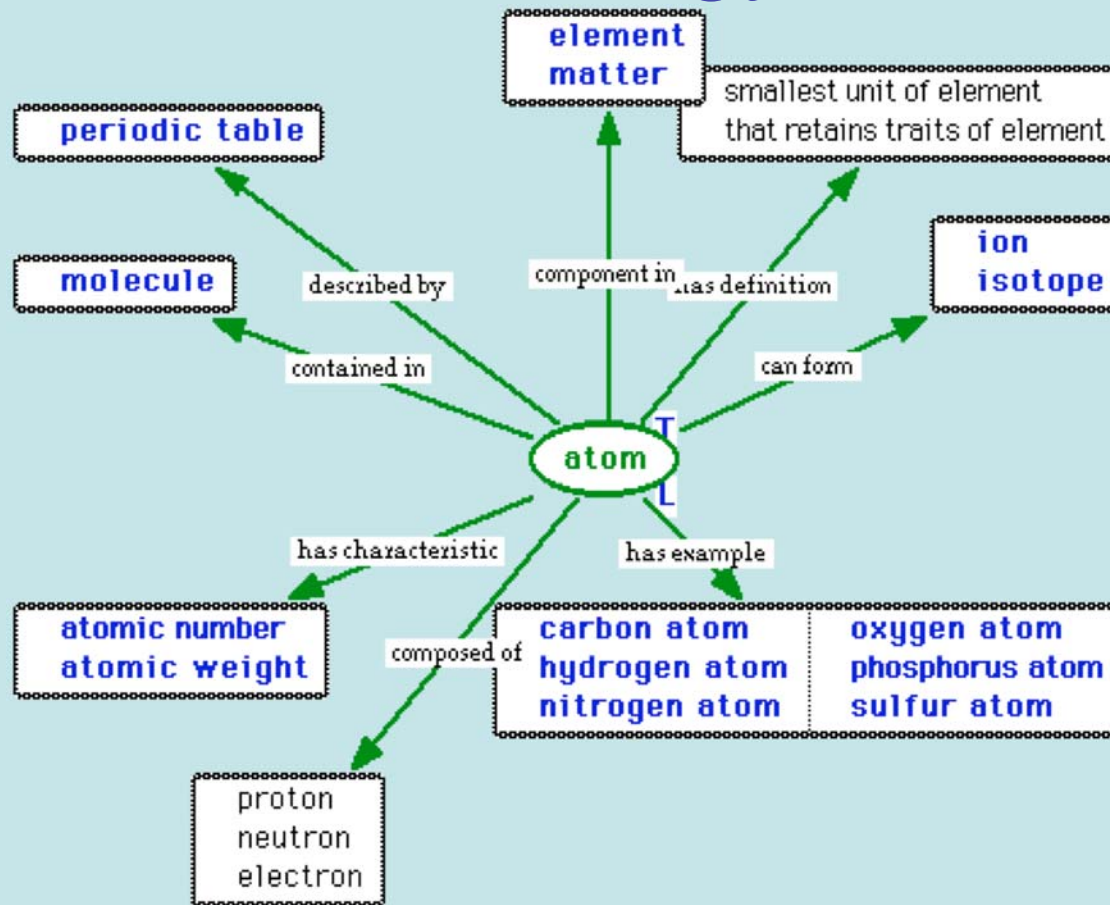


Example SKOS



So every "word" in the thesaurus becomes a uniquely named, Web linkable, concept

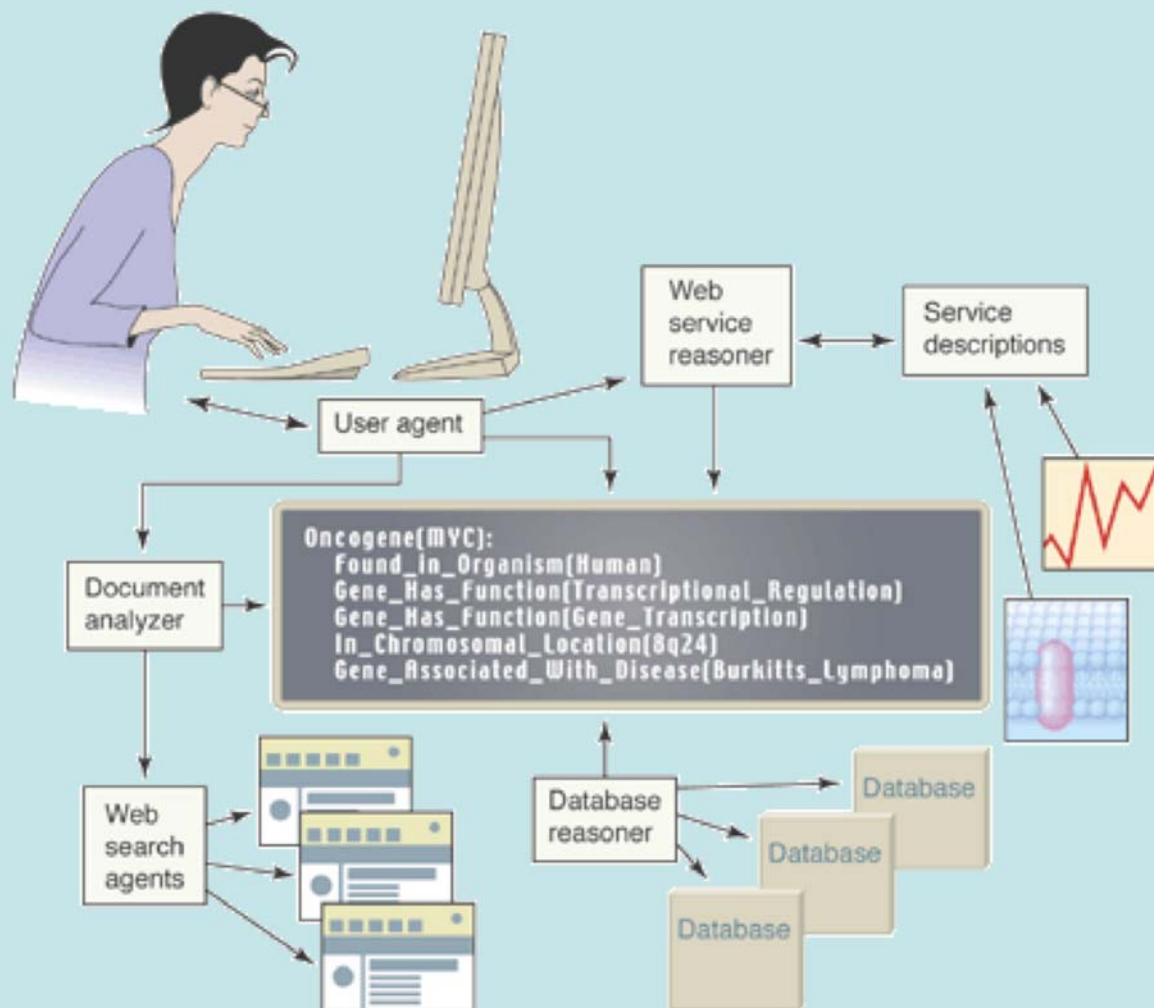
Ontology



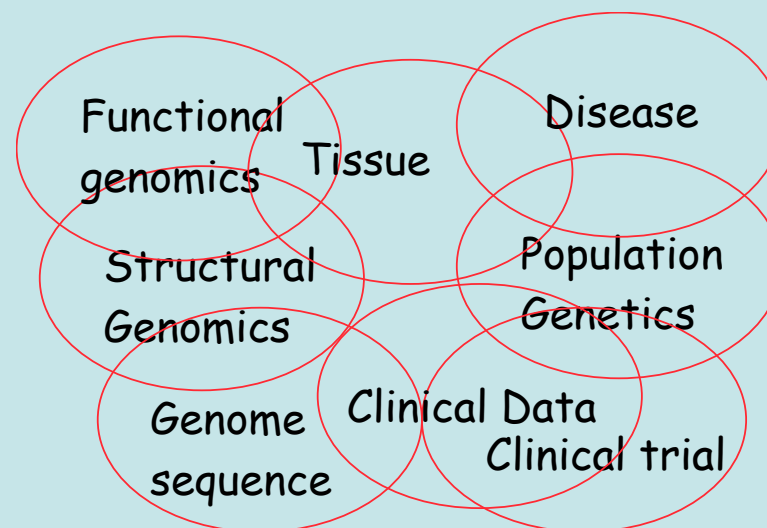
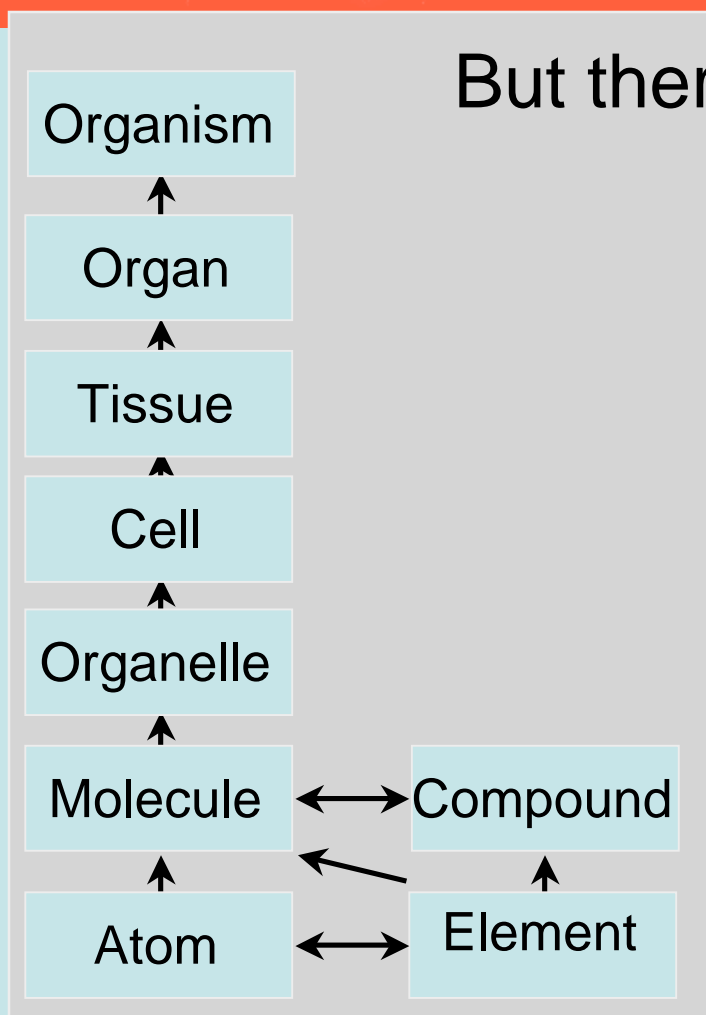
And every "word" in the ontology becomes a uniquely named, Web linkable, concept

Using the links

- These models allow linking of
 - multimedia
 - databases
 - services
 - web
 - Grid?
 - meta-data repos
- Or any other Web resource!



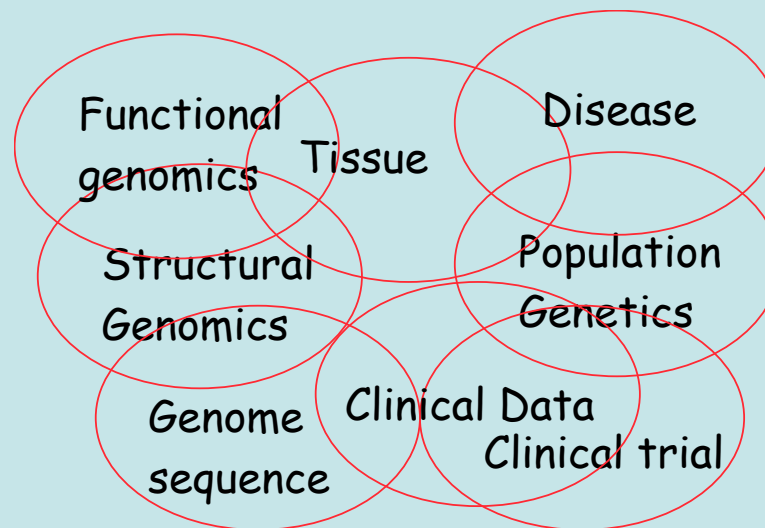
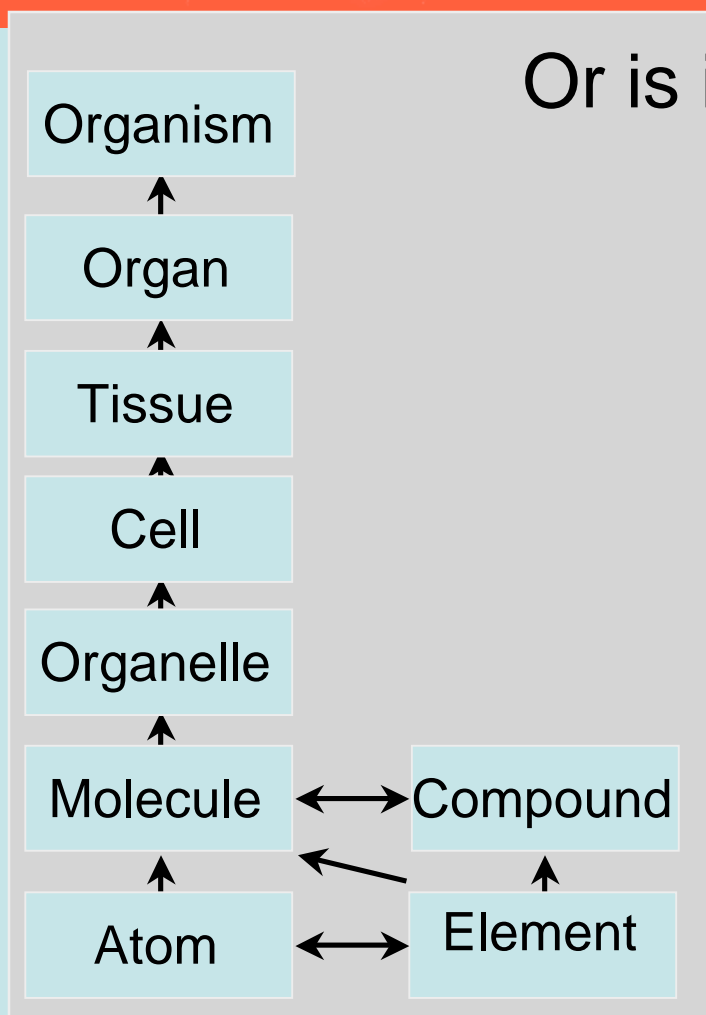
But there's a problem



(Genome World - from Goble, 01)

"... countries separated by a common language"
 -- (Shaw 1942 after Wilde, 1887)

Or is it a feature?

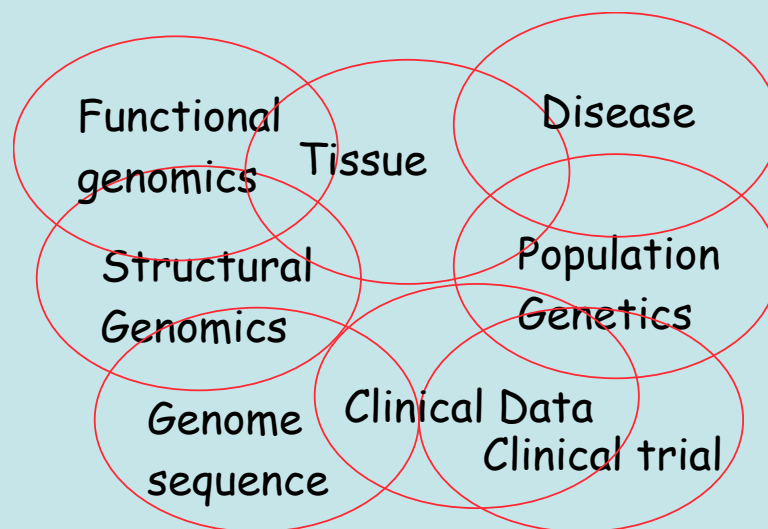


(Genome World - from Goble, 01)

"... countries separated by a common language"
 -- (Shaw 1942 after Wilde, 1887)

From Monolithic to modular

- Curated Groups of ontologies that "overlap"
 - Consistency where needed
 - Shared terms
 - Partial mappings okay
 - Higher level terms
 - Domain specific upper ontologies
 - "backbone" ontologies
 - Existing large-scale consistent domains (c.f. NCI metathesaurus)
 - Existing Thesauri
 - Global consistency not guaranteed
 - And highly over-rated



(Genome World - from Goble, 01)

- **Community model** for development and support
 - C.f. Open Biological Ontologies Consortium
(<http://obo.sourceforge.net/>)
 - Gene Ontology Consortium

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE rdf:RDF [
  <!ENTITY feleuk.owl "http://www.mindswap.org/ontologies/feleuk.owl">
  <!ENTITY owl "http://www.w3.org/2002/07/owl#">
  <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#">
  <!ENTITY NCI "http://www.ncibi.nih.gov/NCIT/NCIT.owl#">
  <!ENTITY CYC="http://www.cyc.com/2004/06/04/cyc#">
]>
<rdf:RDF xml:base="&feleuk.owl;"
  xmlns:owl="&owl;"
  xmlns:rdf="&rdf;"
  xmlns:rdfs="&rdfs;"
  xmlns:NCI="&NCI;"
  xmlns:CYC="&CYC;">
```

Linking is power!

```
<owl:Ontology rdf:about=""
  rdfs:label="Feline Leukemia"
  owl:versionInfo="Feline Leuk 1.0"/>
```

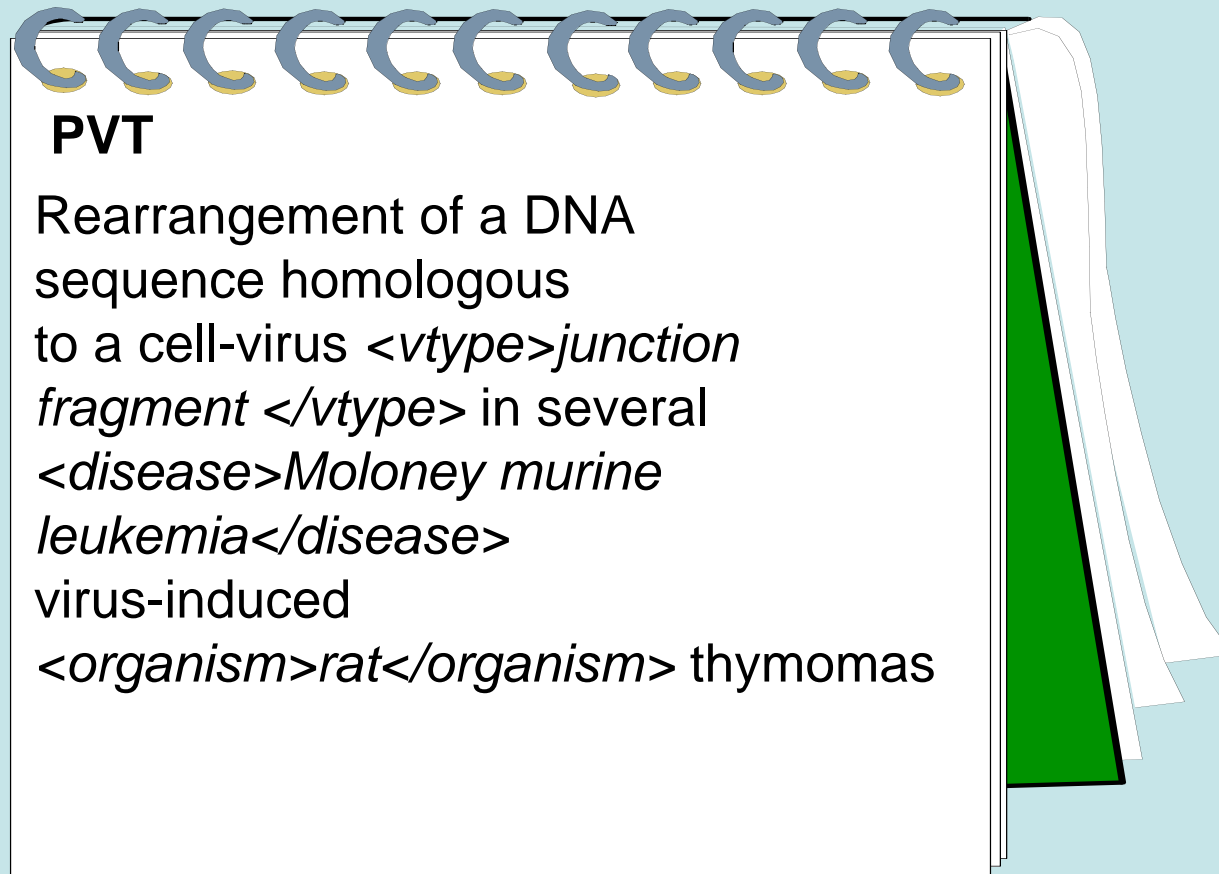
Link to 45000
terms at NCI

```
<owl:Class rdf:about="#Feline-Leukemia">
  <rdfs:subClassOf rdf:resource="NCI:Leukemia"/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:allValuesFrom rdf:resource="CYC:cat"/>
      <owl:onProperty rdf:resource="#NCI:diseased-organism"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

Link to 47000 (Open)CYC terms

```
</rdf:RDF>
```

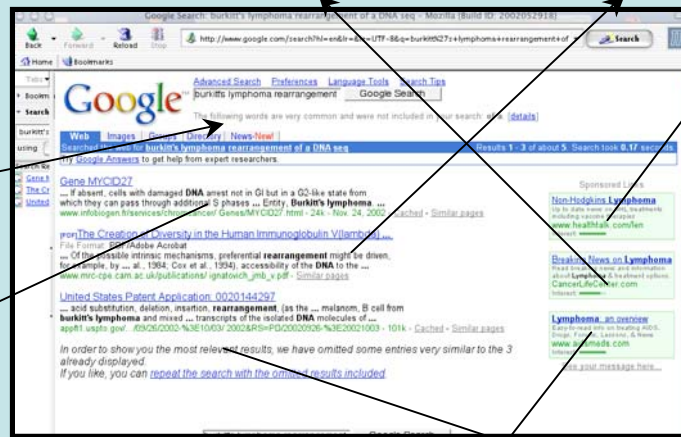
Adding XML



Adding tags, using a document-oriented schema

But that isn't "semantics"

Burkitt's Lymphoma



PVT

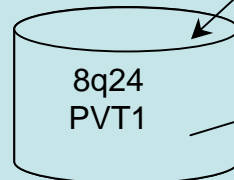
Rearrangement of a DNA sequence homologous to a <cell-type>cell-virus junction fragment </cell-type>in several <disease>Moloney murine leukemia</disease> virus-induced <organism>rat</organism> thymomas

PubMed

Semantic Web

Burkitt's Lymphoma

Oncogene(MYC):
Found_In_Organism(Human).
Gene_Has_Function(Transcriptional_Regulation).
Gene_Has_Function(Gene_Transcription).
In_Chromosomal_Location(8q24).
Gene_Associated_With_Disease(Burkitts_Lymphoma).



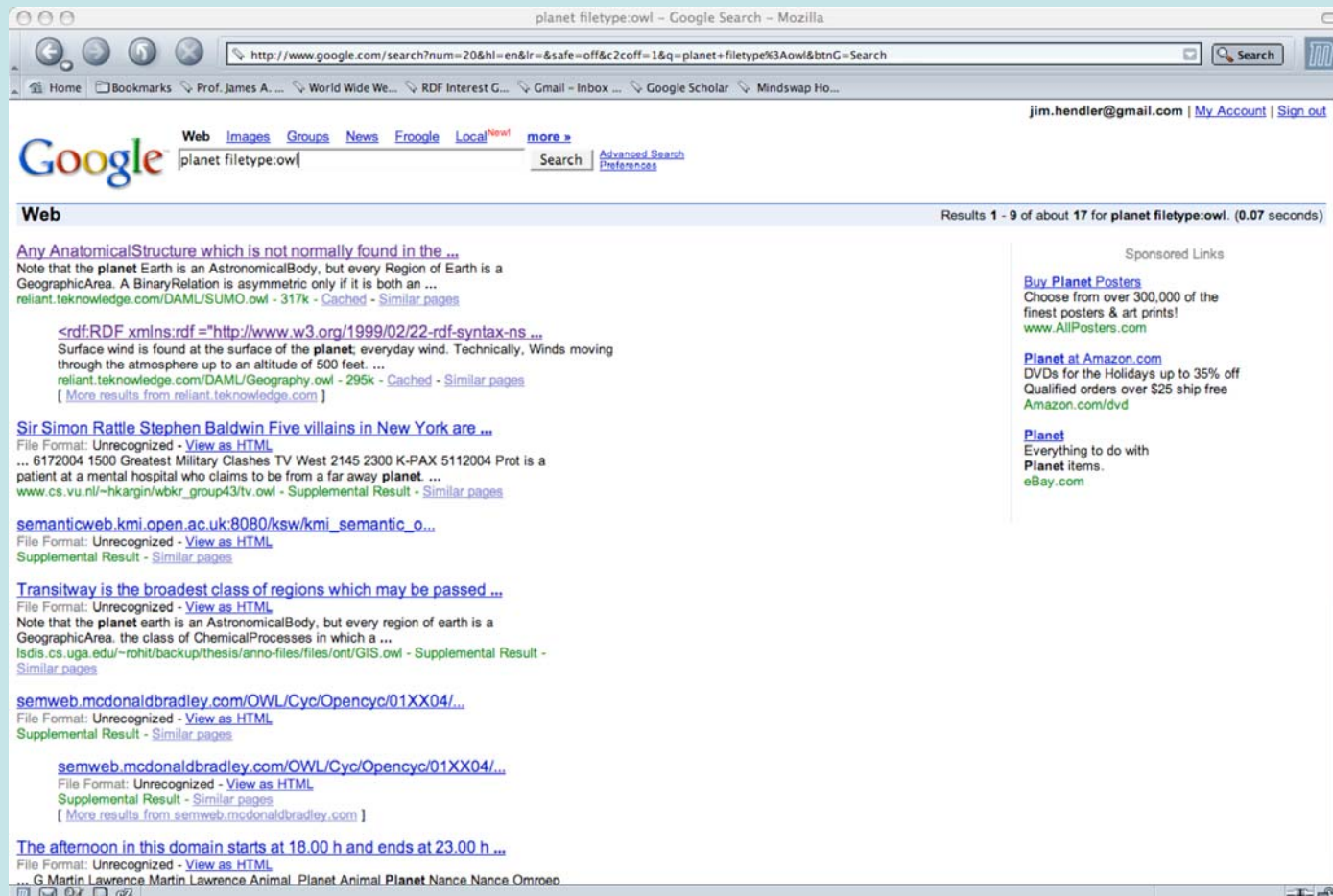
PVT

Rearrangement of a DNA sequence homologous to a cell-virus junction fragment in several Moloney murine leukemia virus-induced rat thymomas

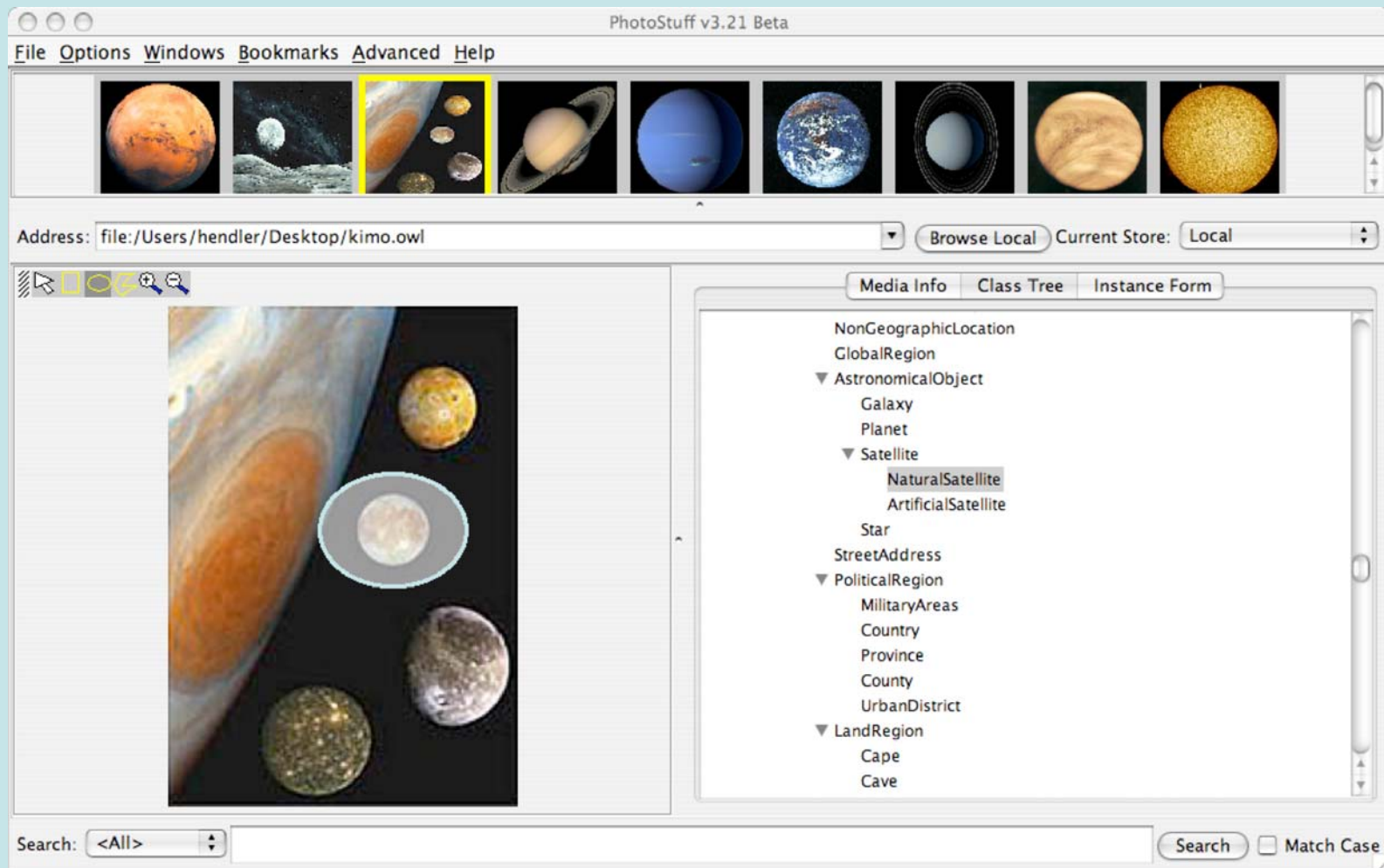
PubMed

Tools

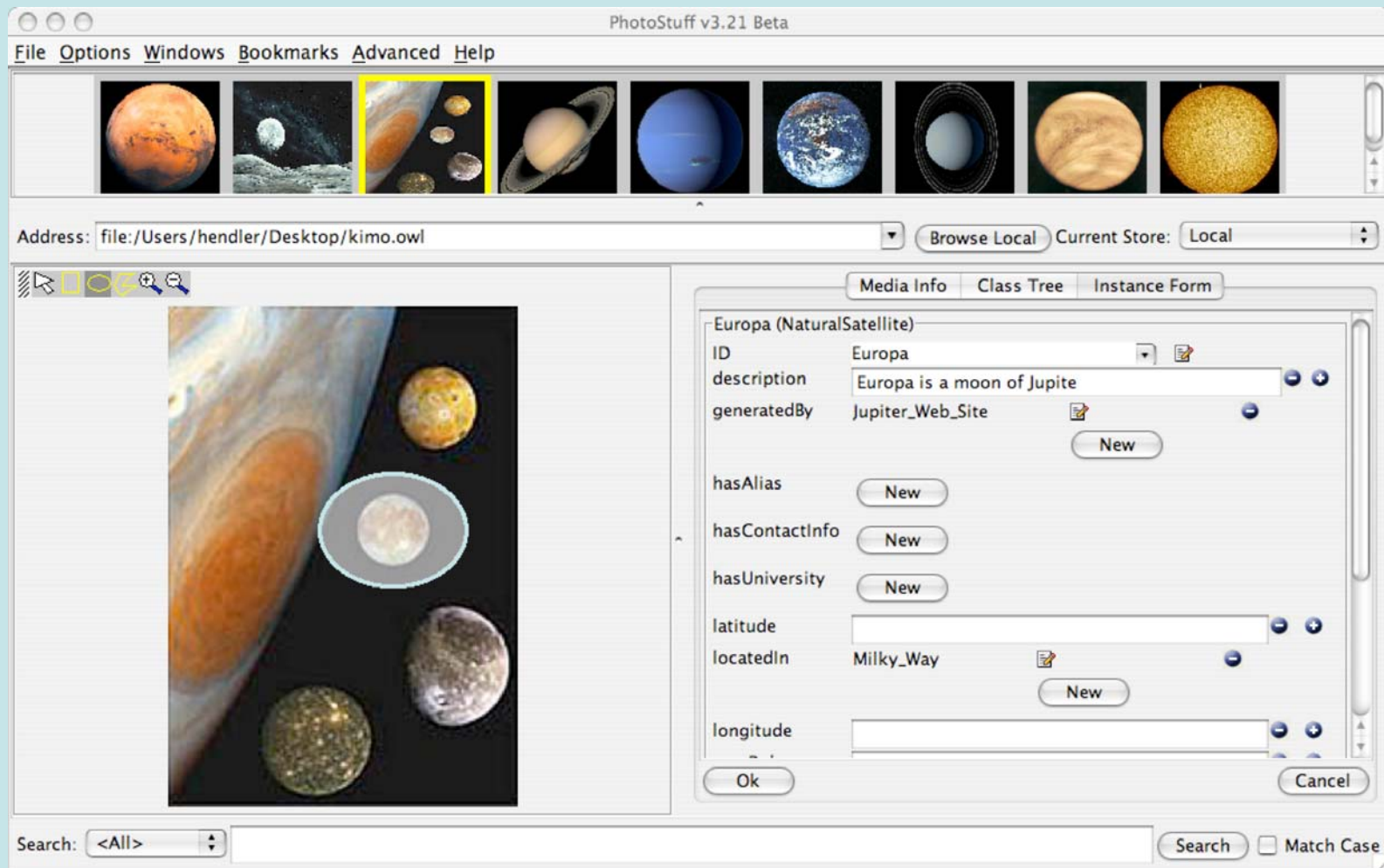
Is it hard to find ontologies?



Use Google®

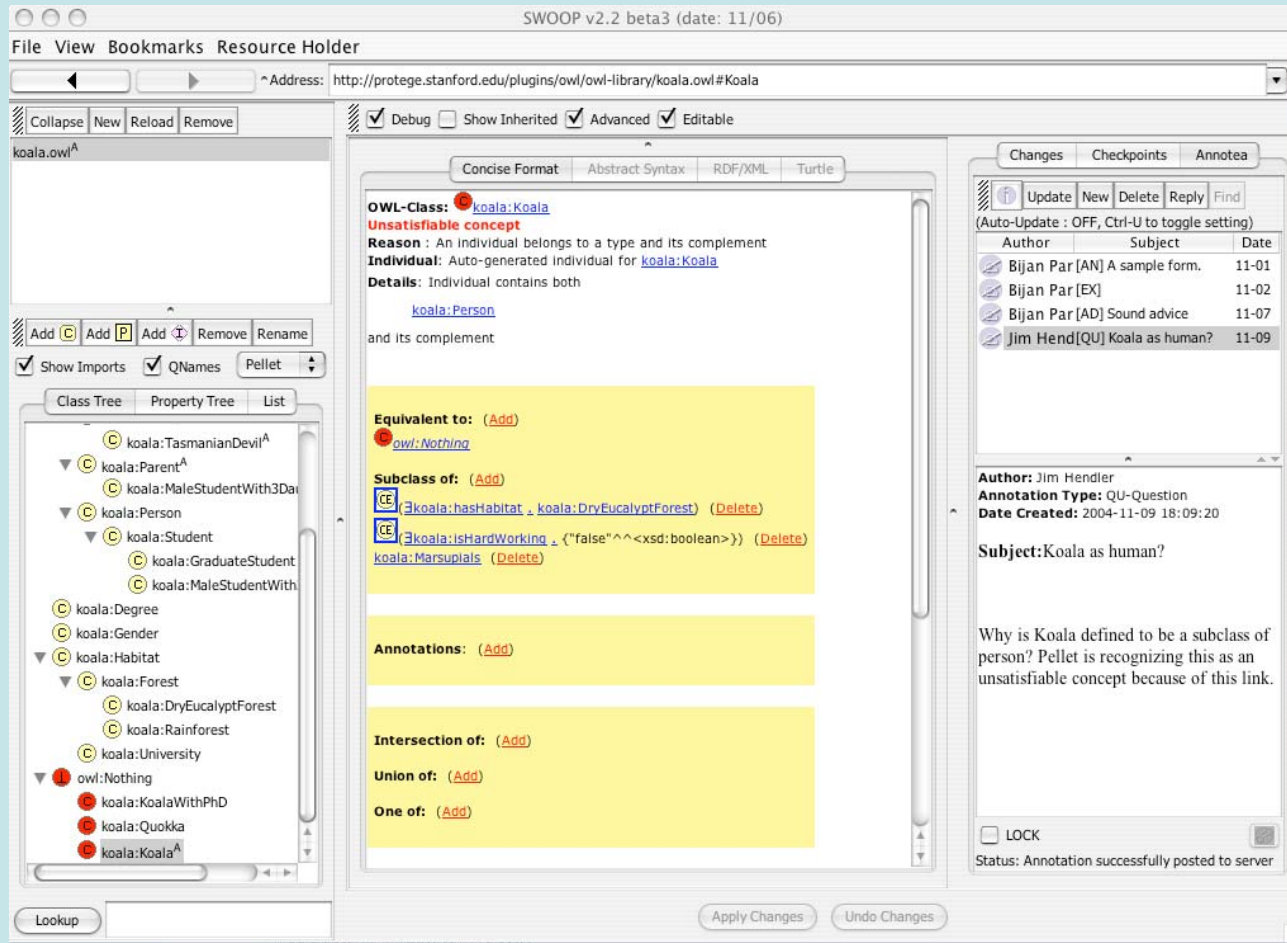


<http://www.mindswap.org/2003/PhotoStuff/>



<http://www.mindswap.org/2003/PhotoStuff/>

SWOOP: OWL ontology tool



SWOOP v2.2 beta3 (date: 11/06)

File View Bookmarks Resource Holder

Address: <http://protege.stanford.edu/plugins/owl/owl-library/koala.owl#Koala>

☒ Debug ☐ Show Inherited ☒ Advanced ☒ Editable

Concise Format Abstract Syntax RDF/XML Turtle

OWL-Class: [koala:Koala](#)
Unsatisfiable concept
Reason : An individual belongs to a type and its complement
Individual: Auto-generated individual for [koala:Koala](#)
Details: Individual contains both
[koala:Person](#)
 and its complement

Equivalent to: [\(Add\)](#)
[owl:Nothing](#)

Subclass of: [\(Add\)](#)
[koala:hasHabitat](#) [koala:DryEucalyptForest](#) [\(Delete\)](#)
[koala:isHardWorking](#) [{"false"^^<xsd:boolean>}](#) [\(Delete\)](#)
[koala:Marsupials](#) [\(Delete\)](#)

Annotations: [\(Add\)](#)

Intersection of: [\(Add\)](#)

Union of: [\(Add\)](#)

One of: [\(Add\)](#)

Changes Checkpoints Annotations

Update New Delete Reply Find

(Auto-Update : OFF, Ctrl-U to toggle setting)

Author	Subject	Date
Bijan Par [AN]	A sample form.	11-01
Bijan Par [EX]		11-02
Bijan Par [AD]	Sound advice	11-07
Jim Hend [QU]	Koala as human?	11-09

Author: Jim Hendler
Annotation Type: QU-Question
Date Created: 2004-11-09 18:09:20

Subject: Koala as human?

Why is Koala defined to be a subclass of person? Pellet is recognizing this as an unsatisfiable concept because of this link.

☐ LOCK
 Status: Annotation successfully posted to server

Apply Changes Undo Changes

Collaboration support

SWOOP v2.2 beta3 (date: 11/06)

File View Bookmarks Resource Holder

Address: <http://www.mindswap.org/2004/owl/funding#GovernmentGrant>

Debug ☐ Show Inherited ☒ Advanced ☐ Editable

Concise Format Abstract Syntax RDF/XML Turtle

Ontology: owl:funding

OWL-Class: funding:Grant

Annotations:
rdfs:label "Grant"

Subclass of:
funding:Funding

Superclass of:
funding:GovernmentGrant
funding:BusinessGrant

Domain of:
funding:funds
funding:name
funding:fundedBy
funding:logo
funding:source
funding:destination
funding:homepage

Domain of:
active-portal-ontology-latest:has-telephone-number
active-portal-ontology-latest:has-funding-source
active-portal-ontology-latest:has-grant-reference
active-portal-ontology-latest:has-author
akt-support-ontology-latest:has-pretty-name
akt-support-ontology-latest:has-magnitude
akt-support-ontology-latest:has-pretty-value
akt-support-ontology-latest:has-variant-name
active-portal-ontology-latest:has-web-address

Range of:
active-portal-ontology-latest:confers-award
active-portal-ontology-latest:uses-resource
active-portal-ontology-latest:produces-output
active-portal-ontology-latest:has-funding

Remove this Entity Remove this Entity

Class Tree Property Tree List

Show All

- funding:BusinessContract
- funding:BusinessFunding
- funding:BusinessGrant
- funding:Contract
- funding:destination
- funding:fundedBy
- funding:Funder
- funding:Funding
- funding:funds
- funding:Government
- funding:GovernmentContract
- funding:GovernmentGrant
- funding:Grant
- funding:homepage
- funding:logo
- funding:name
- funding:Organization
- rdfs:label
- funding:source

Lookup grant

Semantic Web Services

Service composition

File Options

Select a category: **SensorService (14)**

Location

Latitude: in the range

Longitude: greater than

Altitude: equals

Quality: Excellent

Advanced...

Diagram illustrating the service composition process:

```

graph TD
    SoundIntensity[SoundIntensity] --> RMSCalculator[RMS Calculator]
    RMSCalculator --> InputWaveFile[InputWaveFile]
    InputWaveFile --> SoundOutput[Sound Output]
    SoundOutput --> FIRFilter[FIR Filter]
    FIRFilter --> WindowType[WindowType]
    FIRFilter --> LowerFreqLimit[Lower FreqLi mit]
    FIRFilter --> UpperFreqLimit[Upper FreqLi mit]
    FIRFilter --> SoundInput[SoundInput]
    WindowType --> UserInput1[User Input]
    LowerFreqLimit --> UserInput2[User Input]
    UpperFreqLimit --> UserInput3[User Input]
    SoundInput --> Services[Services 4/5]
    UserInput1 --> Rectangular[Rectangular]
    UserInput2 --> InputBox1[ ]
    UserInput3 --> InputBox2[ ]
    Services --> ServicesList[Acoustic Sensor 1, Acoustic Sensor 2, Acoustic Sensor 3, Acoustic Sensor 4]
    
```

Run

*Information management capabilities
Discovery, Filtering, Composition*

Semantics and services

input xsd:complex="oncogene"

```
<?xml version='1.0'?>
<xsd:schema xmlns="http://www.w3.org/2001/XMLSchema"
  targetNamespace="urn:GoogleSearch" >
  <xsd:complexType name="GoogleSearchResult" >
    <xsd:all>
      <xsd:element name="documentFiltering" type="xsd:boolean" />
      <xsd:element name="searchComments" type="xsd:string" />
      <xsd:element name="estimatedTotalResultCount" type="xsd:int" />
      <xsd:element name="estimateIsExact" type="xsd:boolean" />
      <xsd:element name="resultElements" type="typens:ResultElementAr" />
      <xsd:element name="searchQuery" type="xsd:string" />
      <xsd:element name="startIndex" type="xsd:int" />
      <xsd:element name="endIndex" type="xsd:int" />
      <xsd:element name="searchTips" type="xsd:string" />
      <xsd:element name="directoryCategories" type="typens:DirectoryCa" />
      <xsd:element name="searchTime" type="xsd:double" />
    </xsd:all>
  </xsd:complexType>
</xsd:schema>
</types>
<message name="doGoogleSearch">
  <part name="key" type="xsd:string" />
  <part name="q" type="xsd:string" />
  <part name="start" type="xsd:int" />
  <part name="maxResults" type="xsd:int" />
  <part name="filter" type="xsd:boolean" />
  <part name="restrict" type="xsd:string" />
  <part name="safeSearch" type="xsd:boolean" />
  <part name="lr" type="xsd:string" />
  <part name="ie" type="xsd:string" />
  <part name="oe" type="xsd:string" />
</message>
<message name="doGoogleSearchResponse">
  <part name="return" type="typens:GoogleSearchResult" />
</message>
<operation name="doGoogleSearch">
  <input message="typens:doGoogleSearch" />
  <output message="typens:doGoogleSearchResponse" />
</operation>
</definitions>
```

Oncogene(MYC):

Found_In_Organism(Human).

Gene_Has_Function(Transcriptional_Regulation).

Gene_Has_Function(Gene_Transcription).

In_Chromosomal_Location(8q24).

Gene_Associated_With_Disease(Burkitts_Lymphoma).

output xsd:complex="RiskType"

```
<owl:Class rdf:about="http://annotation.semanticweb.org/iswc/iswc.daml#RiskIndicator">
```

```
<rdf:subClassOf>
```

```
<owl:Restriction>
```

```
<owl:onProperty rdf:resource="http://annotation.semanticweb.org/iswc/iswc.daml#name"/>
```

```
<owl:allValuesFrom rdf:resource="http://www.w3.org/2000/10/XMLSchema#string"/>
```

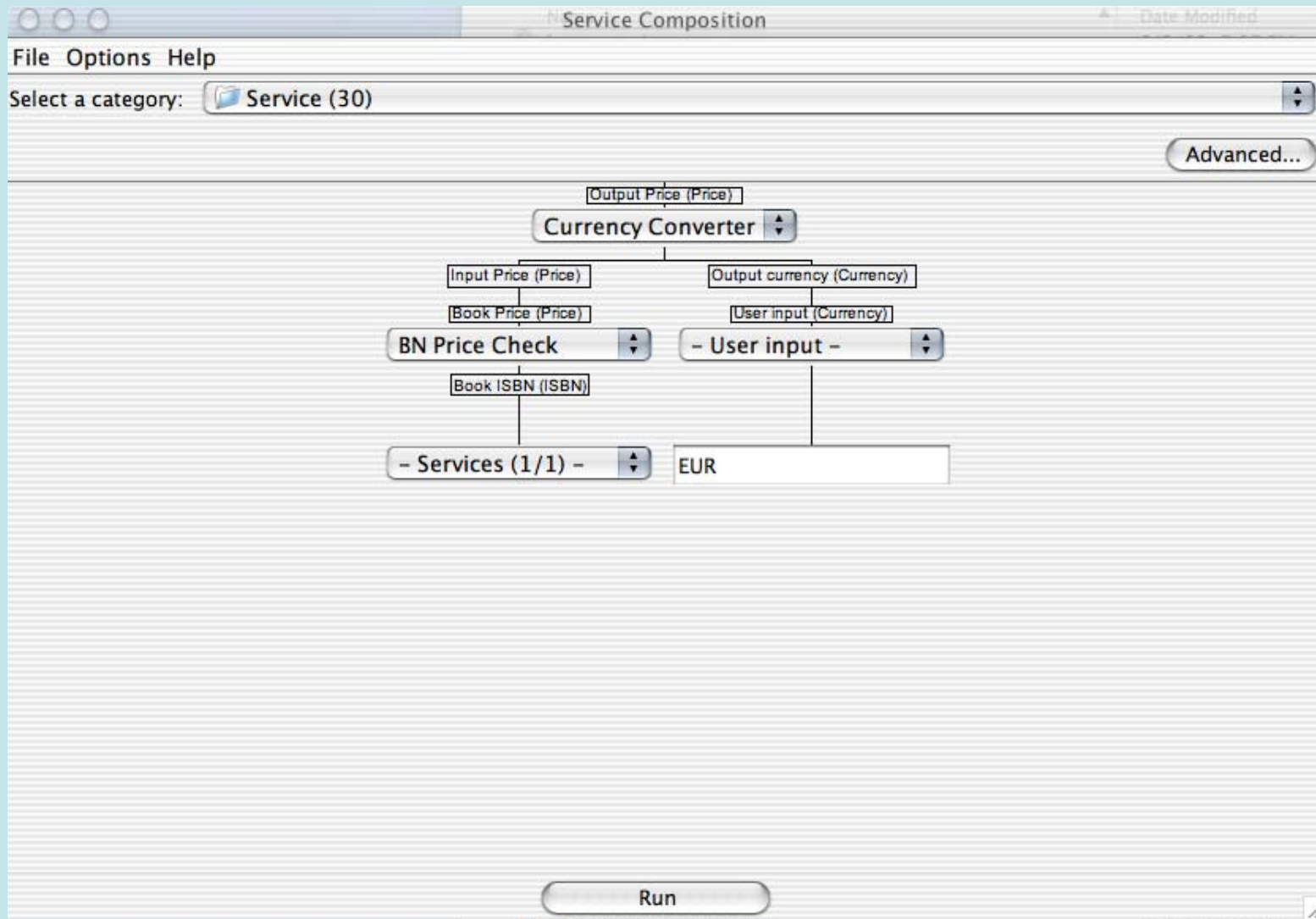
```
</owl:Restriction>
```

```
</rdf:subClassOf>
```

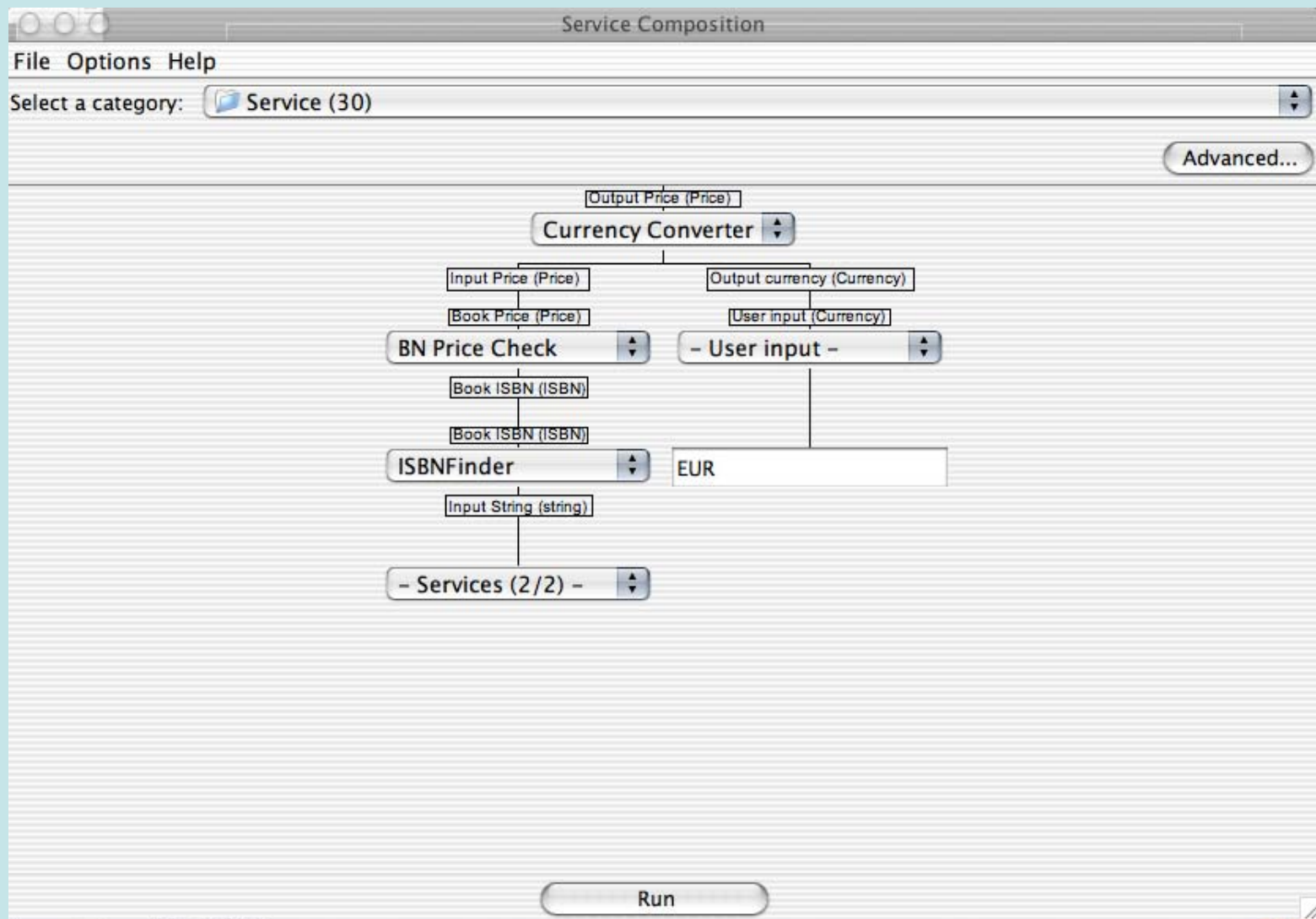
```
</:Class>
```

OWL-S, WSDL-S, WSDL2RDF

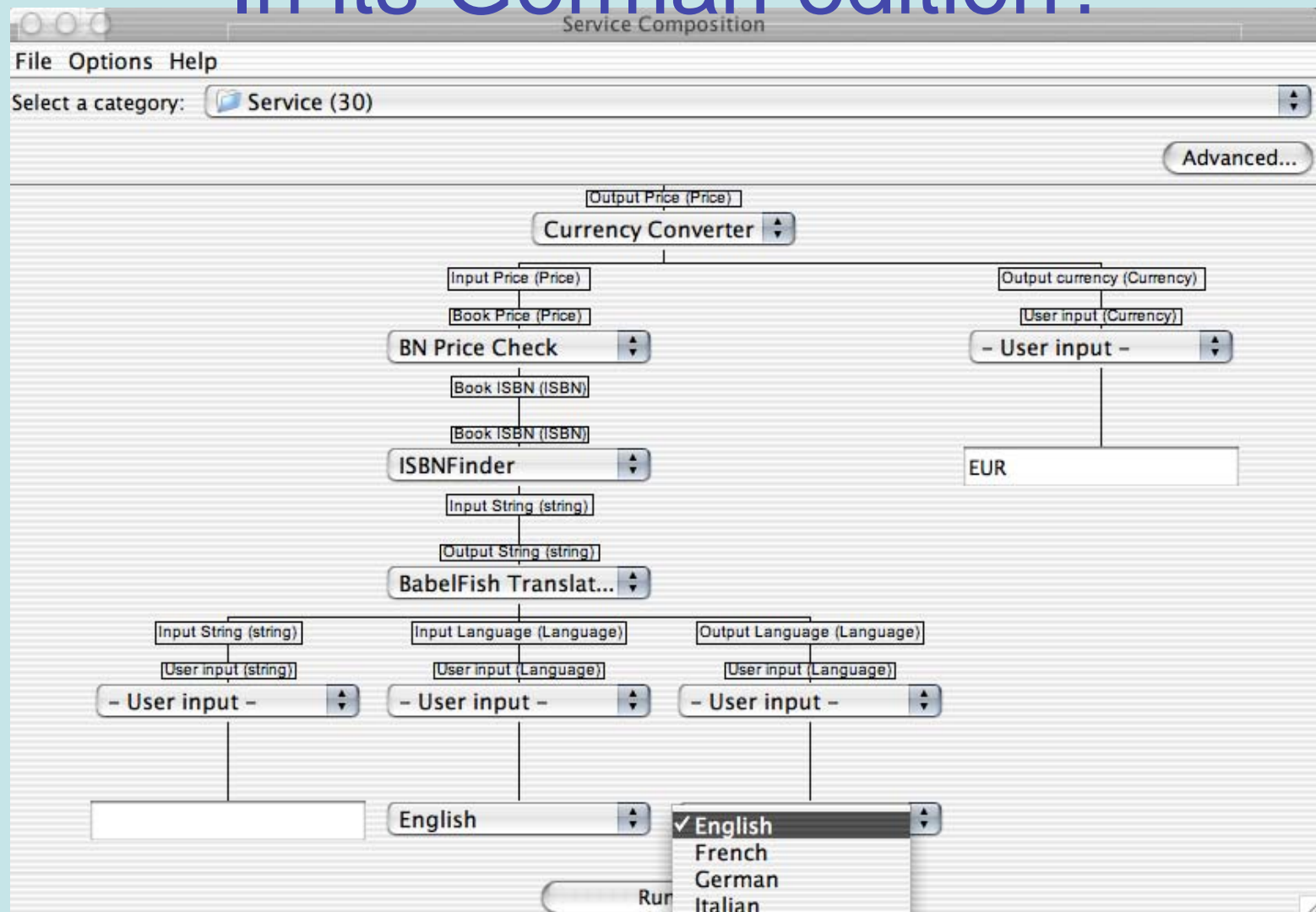
Get a B&N price (In Euros)

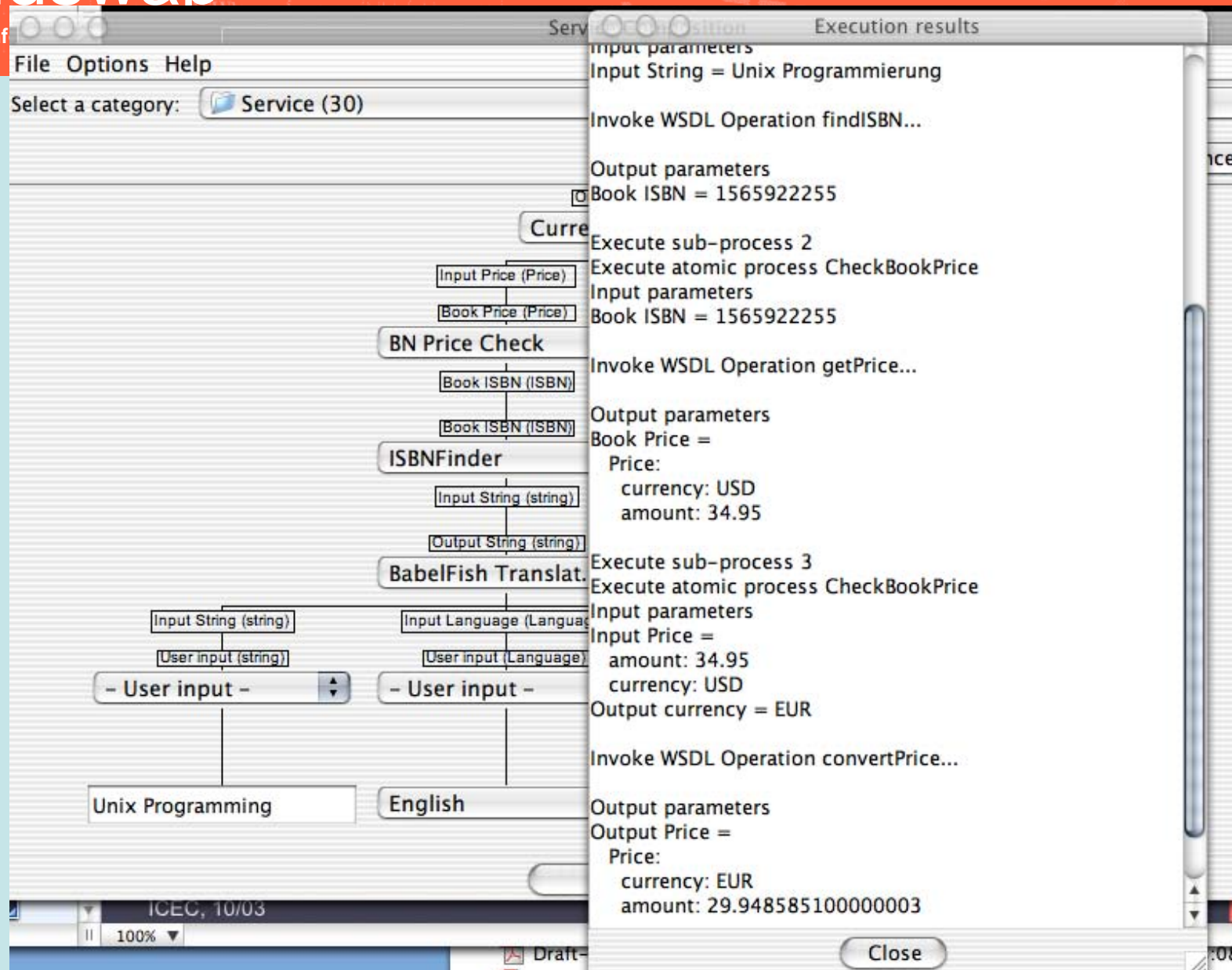


Of a particular book



In its German edition?

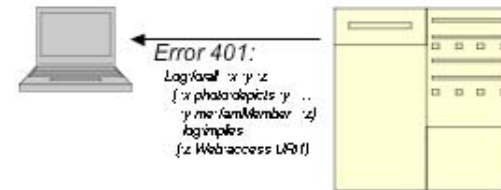




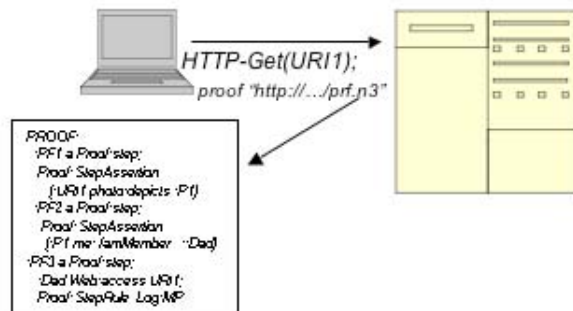
Policy Aware WEB



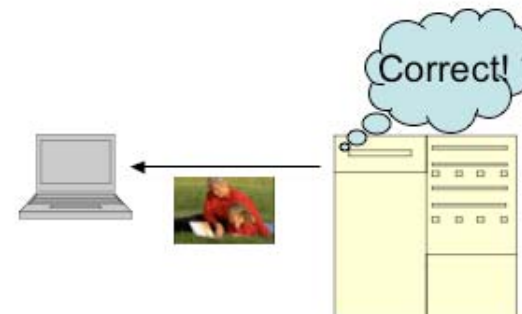
(A) User requests a resource.



(B) 401 error provides access rules.



(C) Proof is generated and pointer is sent in new HTTP-Get request.



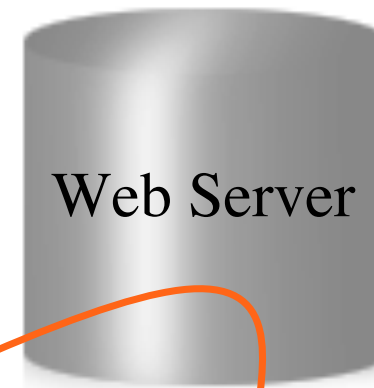
(D) Proof is checked, and confirmed, and the transaction succeeds.

PAW demo



Use case:

A Web browser requests the home page for a girl scout troop and is given it by a Web server.



Web Server

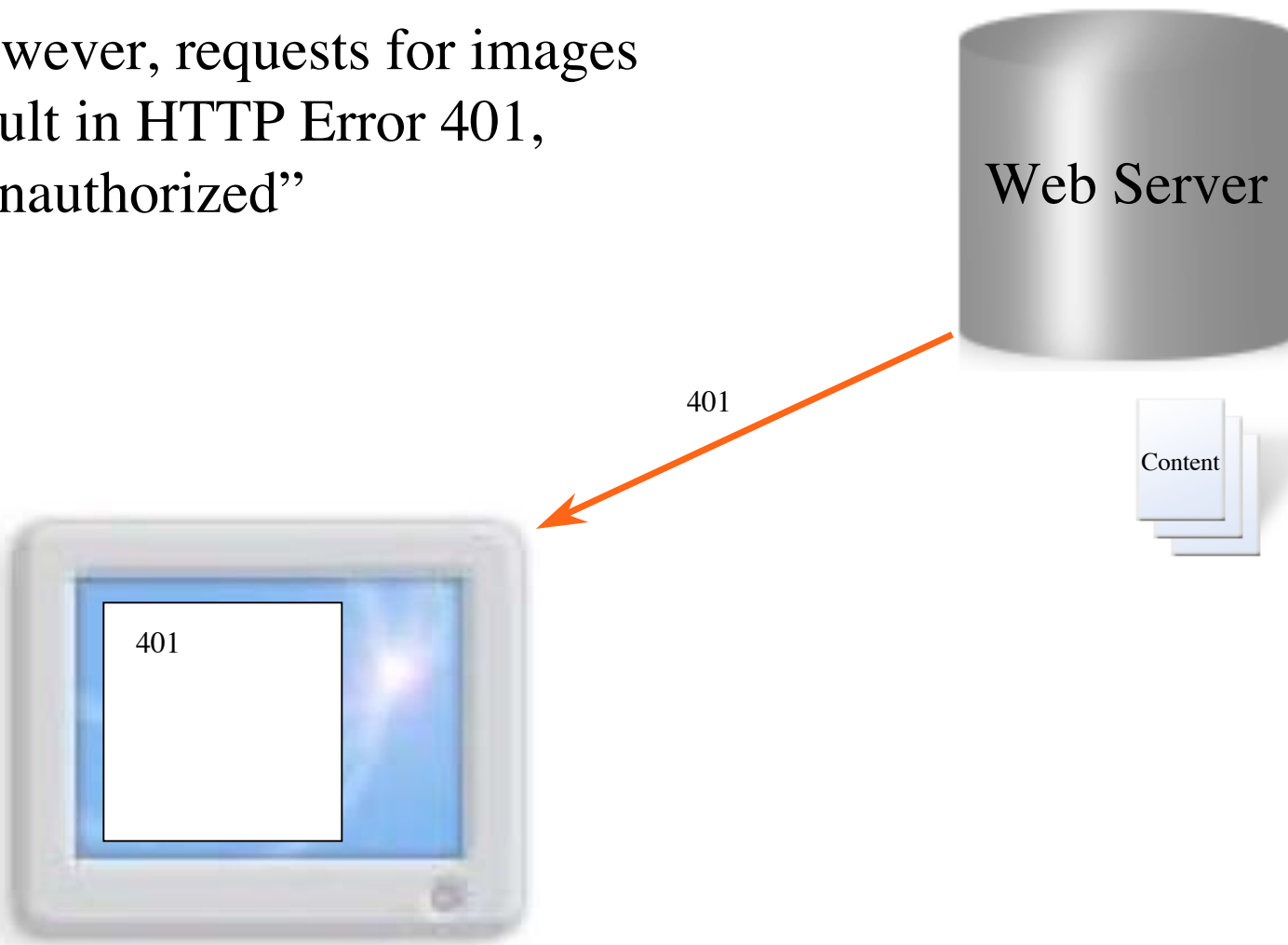


Content

Demo

Harvard IIC, 12/2005

However, requests for images
result in HTTP Error 401,
“Unauthorized”



The 401 “Unauthorized” response
has been modified to provide a
URL to a policy:

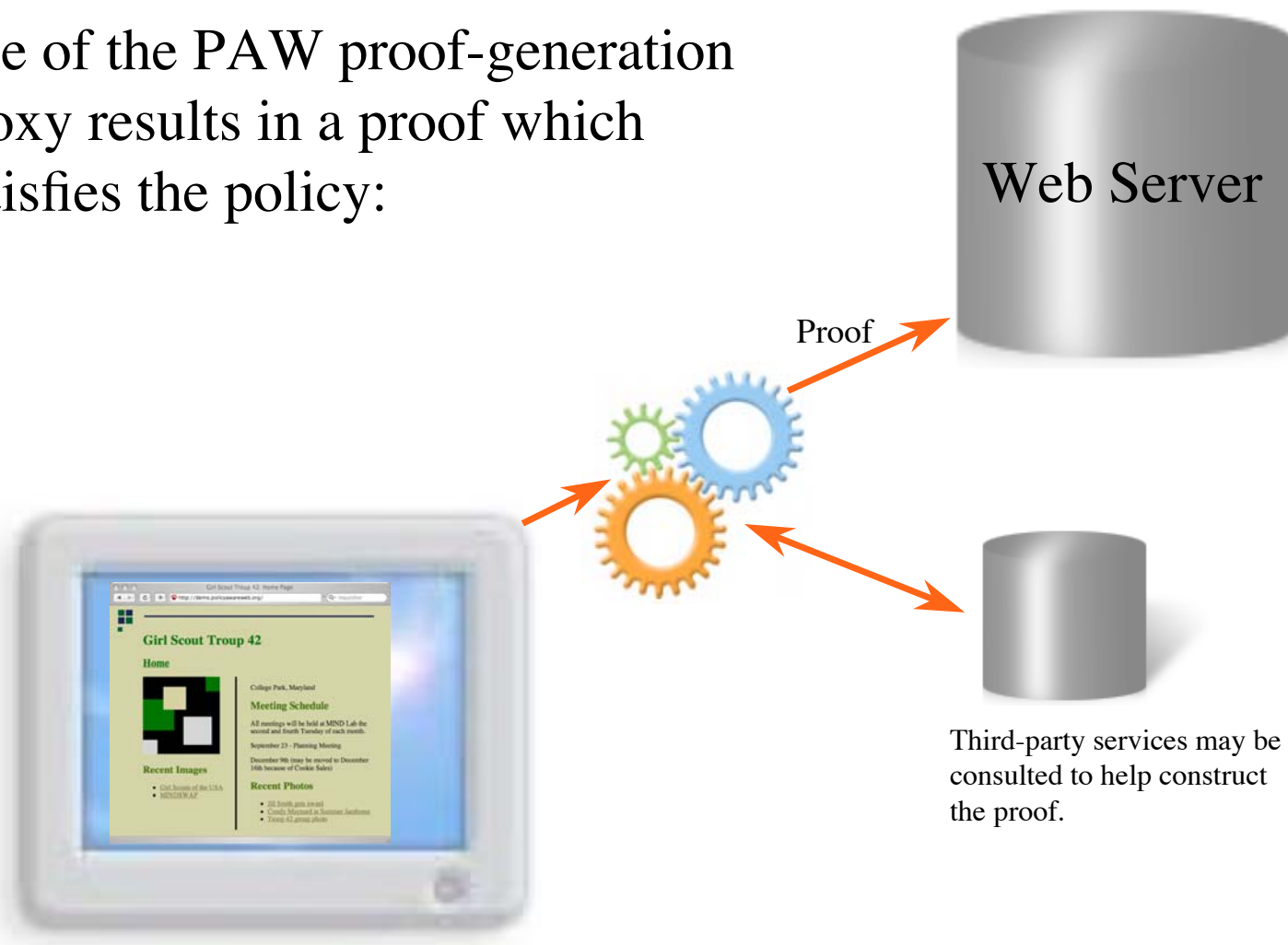
```
HTTP/1.1 401 Not authorized
Date: Sat, 03 Dec 2005 15:32:18 GMT
Server: TwistedWeb/2.0.1
Policy: http://groups.csail.mit.edu/dig/2005/09/rein/examples/troop42-policy.n3
Content-type: text/html; charset=UTF-8
Connection: close
10:32:20 ERROR 401: Not authorized.
```


Policies use linked rules

- Example policies
 - Photos taken at meetings of the troop can be shared with any current member of the troop.
 - Photos taken at a jamboree can be shared with anyone in the troop or with anyone who attended the jamboree.
 - Photos of any girl in the troop can be shared with the world if that girl's parent has given permission (under construction)

```
{ REQ a rein:Request.  
  REQ rein:resource PHOTO.  
  ?F a TroopStuff; log:includes  
    { PHOTO a t:Photo; t:location LOC.  
      LOC a t:Meeting }.  
  
  REQ rein:requester WHO.  
  WHO session:secret ?S.  
  ?S crypto:md5 TXT.  
  
  ?F a TroopStuff; log:includes  
    { [] t:member [ is foaf:maker of PG ].  
      LOC t:attendee [ is foaf:maker of PG ] }.  
  PG log:semantics [ log:includes  
    { PG foaf:maker [ session:hexdigest TXT ] }  
  ].  
  
} => { WHO http:can-get PHOTO }.
```

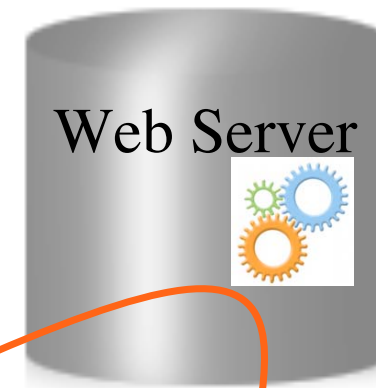
Use of the PAW proof-generation proxy results in a proof which satisfies the policy:



The proxy:

1. Uses Rein, a policy engine, to specify rules which match a given policy.
2. The Rein rules are run in Cwm, a forward-chaining reasoner for the Semantic Web. This generates a proof.

The Web server checks the proof
and serves the content if it is valid.



Web Server



Content



Demo

Harvard IIC, 12/2005

The server:

1. Uses Cwm to validate the proof.
2. Takes action based on validation (serves content or denies).

Current work:

1. Embed Pychinko, a Rete implementation, in Cwm to enhance performance of rules application.
2. Check the proof for satisfaction of a policy rule at the proxy before submission to the server.
3. Make use of multiple distributed authentication systems (instead of holding secrets in the proxy).
4. Associate content with RDF metadata and base policy decisions on the RDF (cf. policy 3)
5. Address issues of eventual integration of the proxy with a Web browser (e.g. cookie storage).

Where were we?

- I would like to expand the vignette in the cyberinfrastructure web site starting with an explanation of why the astronomer was in the coffee shop that morning. You see, she had just pulled an all-nighter doing the work she needed to to get the computation ready to run. First, she had to spend a couple of hours using Google to see if she could **find some programs to use for her simulation models**. She then spent a couple of hours searching through the appendices of papers she found on the open-source physics archives to **find the datasets** that represented the particular galaxies she wanted to explore (and which, of course, cannot be searched for in Google which has no search capability against data). At that point, she had to start chatting with colleagues in Japan (who were just waking up at that time of night), because the datasets she had found were not in the format she needed for the program she needed, so she had to **find a program that she could use to convert formats**. Towards dawn, she had all the components she needed, unfortunately, she was **writing some glue code that would create the workflow she needed to execute the whole thing**. Finally, she called friends at a half dozen computer centers as they came in in the morning so she could **get all the passwords and keys that would be needed so her code could run in the distributed system**.

We can take steps towards there by applying semantic technology, and esp. the Semantic Web, to the practice of science

Social issues

- Training and Curriculum
 - Scientific informatics (not just bioinformatics)
- Culture changes
 - Effect of these things on practice of science
 - + Interdisciplinary
 - Disruptive technology for 200-year-old publication model of science
 - Motivations/rewards
 - ? Can we exploit the motivations that lead to the "Netwatch" column in Science?

Conclusion

- Semantic (Web) technologies have much promise to science
- To get there requires
 - Creating generic technologies for information and knowledge management
 - Developing tools to make these human accessible
 - And helpful
 - Tailoring these tools more specifically to scientists needs
 - Esp. publishing
 - Exploring the ramification of open, distributed information models
 - I.e. PAW vs. Key infrastructure
- Which are key challenges to the use of computer science in science
 - And are, alas, not where most work is currently focused

Not discussed

- Interface issues
 - Don't underestimate scientists
- Roadmap issues
 - Sem Web (and other technologies) viz. current practice *and* future research challenges
 - Inconsistency, scalability, algorithmics, uncertainty, etc.
- Other technologies
 - SW dovetails with Grid (VOs), trust/soc network, probabilistic reasoning, machine learning, ...

Acknowledgements

- My students and staff do all the hard stuff...
 - Details at <http://www.mindswap.org/people>
(Our Semantic Web Portal)
- Made possible by our funders
 - Details at <http://www.mindswap.org/funding>

MIND SWAP

- Maryland Information and Network Dynamics Laboratory,
Semantic Web and Agents Project
 - J. Hendler
 - B. Parsia
 - Jennifer Golbeck
 - Aditya Kalyanpur
 - David Wang
 - Daniel Krech
 - Evren Sirin
 - Ronald Alford
 - Jordan Katz
 - Amy Alford
 - Naiwen Lin
 - Kendall Clark
 - Daniel Hewlett
 - Chris Testa
 - Bin Zhao
 - Aaron Mannes
 - Keith Mantel
 - Sharone Horowitz-Hendler
 - Debbie Heisler
- Corporate Research Partners:
 - **Fujitsu Laboratory of America, College Park**
 - Lockheed Martin Advanced Technology Laboratories
 - Northrup Grumman Corporation
 - SAIC Corp., Kevric Corp, Top Quadrant, Tucana, Semaview
- Govt Funding:
 - NSF, US Army Research Laboratory, DARPA, DoD, NIST