Teraflops Petabytes and Exalinks

Science and the Semantic Web

Professor James Hendler

University of Maryland

http://owl.mindswap.org

Keynotes

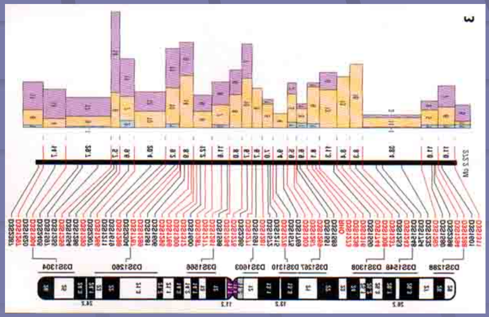
Celebrating the Success of the Grid and its impact on science

“In short, the Grid is tomorrow’s Internet”

MiND

Big iron / Large Pipes

- Tera-flops
  - Despite Moore
    we are compute bound
- Peta-bytes
  - We can collect more
    than we can process
  - And our instruments can
    produce more, faster
- (US) Grid vision focuses on moving lots of data to ever larger computers



What was proposed

- Research investment in US, EU, UK and Japan to create infrastructure for Scientific computing
  - many 100s of millions of dollars spent
  - many high end computers built
  - new Grid infrastructure built
  - scientific databases extended
  - internet2 high bandwidth pipes
  - hundreds of computers bought
- Have we succeeded in “Changing the nature of the practice of science”?

Have we succeeded?

- Most used Scientific Software
  - Publishing Software: Word
  - Presentation Software: Powerpoint
  - Collaboration Software: email
- Successful Scientific infrastructure
  - Data Search: WWW (Google)
  - Paper Search: WWW (Google)
  - Networking: WWW (Client/Server)
- BTW, Direct Connect (Modus 2) has well over a petabyte of info available
  - mostly music and video

ExaLinks

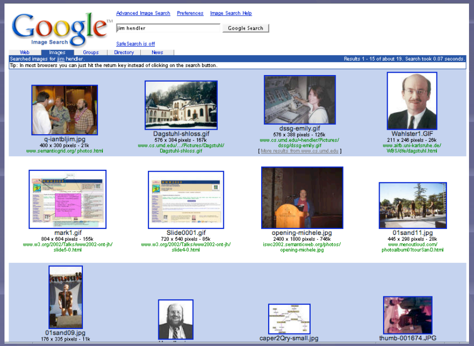
- The Web succeeds by exploiting the “network effect”
  - a graph of >3[illegible]1[illegible] nodes
  - [illegible] pos[illegible]ble [illegible]th[illegible]n [illegible]lie[illegible]arge num[illegible]er [illegible]O(2^3.3e9)
  - Paths of length 4-5
    - approx = 1.2e18

Solved!

MiND

Problems w/the solution

- Doesn't work well for non-text resources
  - minimal image query using text as heuristic
    - bad at video, sound
    - bad at database query, search
    - bad at programs/services
    - surprisingly good at text



- Doesn't work for information not on a single page
  - minimal "query within results" capability
- Path composition managed by humans
  - exa-complexity left to the user

MIND

The World Wide Web (review)

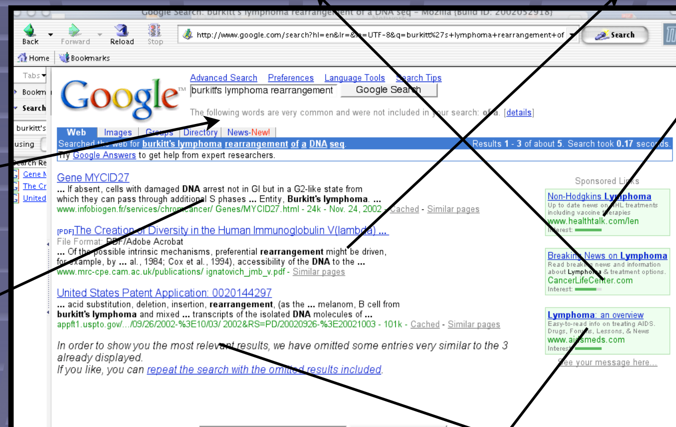
- On the order of 10^8 users
  - Used in ***every*** country on Earth
    - On ***every*** continent (incl. Antarctica), **Mars link is in transit!**
  - A tiny percentage is "trained" in any way
- On the order of 10^{10} indexed web resources (text) in Google etc
  - Essentially Infinite if one includes "dynamic" web pages
- Massively distributed and open
  - ***Anyone*** can play -- To someone on the Web, *you're* the nut
- A set of protocols and languages driven by a strong standards approach
  - Implementation and platform independence crucial
  - World Wide Web Consortium the most prominent
  - If you don't play by the rules, you don't really play
    - Ex. Non-HTML browsers lost big time

What are we doing wrong?

- 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

Network effect: Science must "use the links"

Burkitt's Lymphoma



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

PubMed

Web

Semantic Web

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).

"MODEL"

Burkitt's Lymphoma

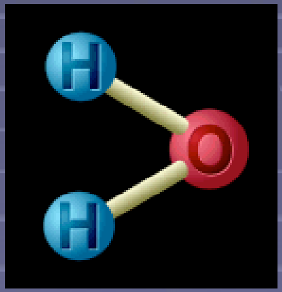
8q24
PVT1

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

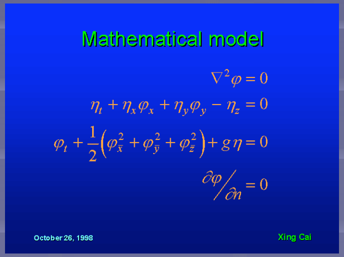
PubMed

A very old idea in new clothes

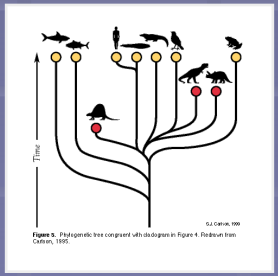
- Scientists communicate by use of models
  - c.f. Physical



  - c.f. Mathematical

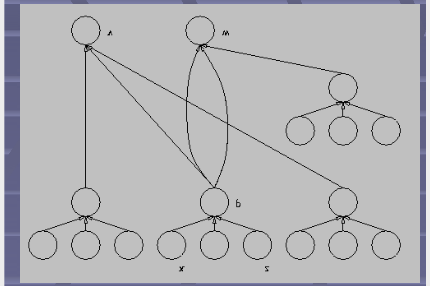


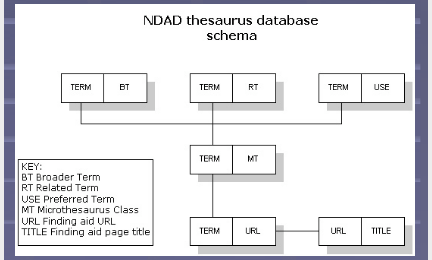
  - c.f. Organizational

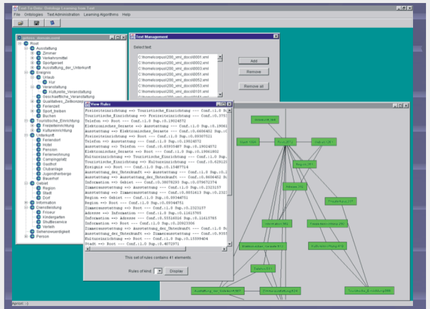


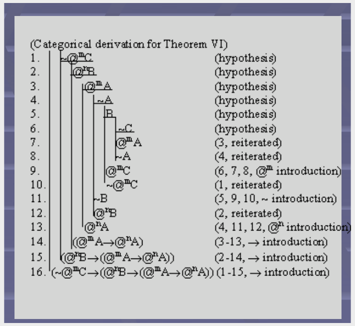
Models expose semantics

Sem Web Modeling









Graph Ontology	Labeled graph Data Dictionary Data Schema Ontology	Graph + limited logic Ontology	. . .	Logic Ontology
RDF	RDF Schema	OWL	. . .	KIF?

All of these languages add semantic modeling primitives to XML - so you can "do this in XML" per se, but it is reinventing the wheel.

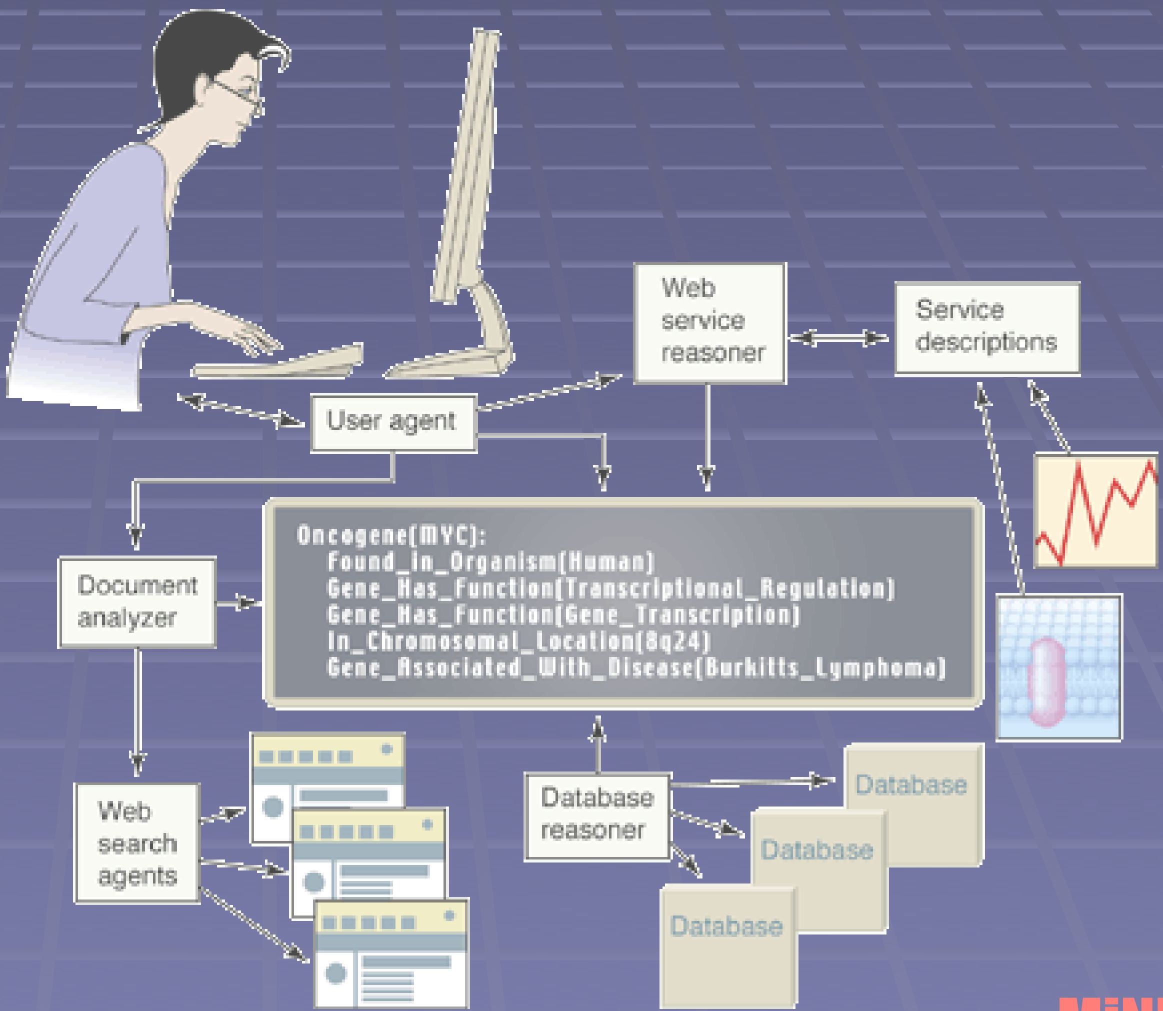
MiND

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
- OWL, the XML Schema of the Semantic Web
  - 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
    - * More

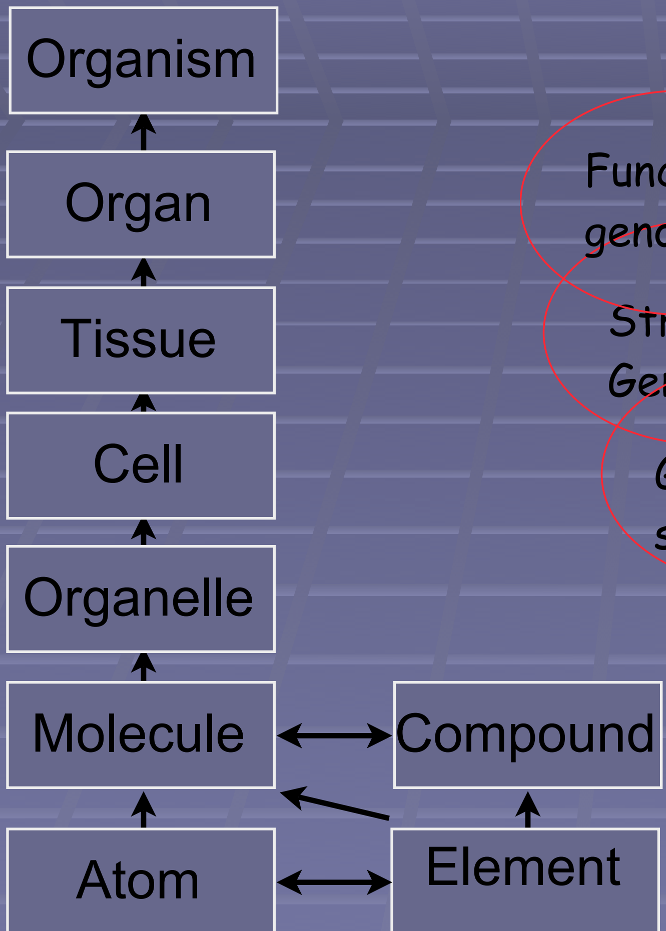
Using the links

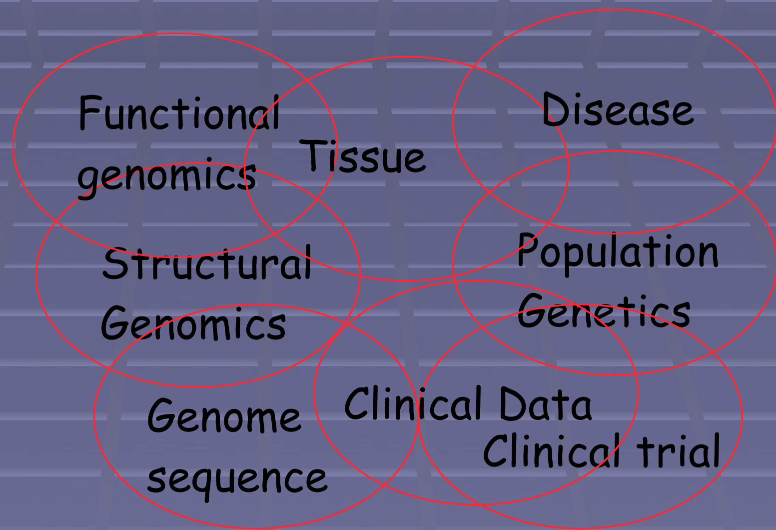
- These models allow linking of
  - multimedia
  - databases
  - services
    - web
    - Grid?
  - meta-data repos
- Or any other Web resource!



MiND

But there's a problem





(Genome World - from Goble, 01)

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

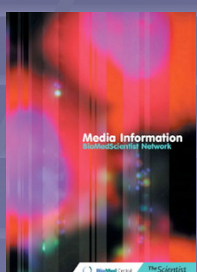
Scientific impact

- 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

Network effect

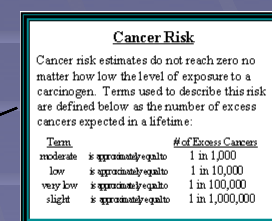
- The models can also link to other models
  - partial mappings just fine
  - this creates a web of models (semantics) much like the current web is a web of texts
- Network effect as mappings provide links to linked resources

BioMedCentral Article



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).

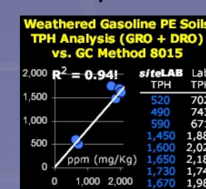
NCI Cancer Ontology (OWL)



EPA Vocabulary (RDFS)

<meta>
<classifications>
<classification type="MYC" subtype="old_arx_id">bcr-2-1-059</classification>
</classifications>
</meta>

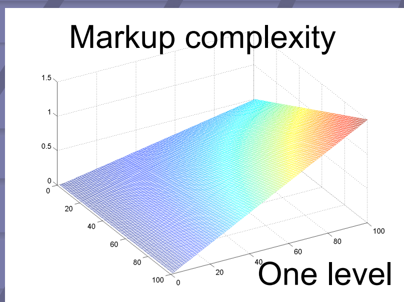
BioMedCentral Metadata (XML)

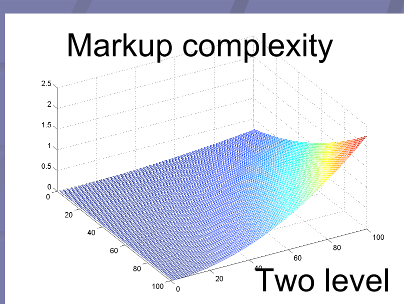


EPA data set (XML)

MIND

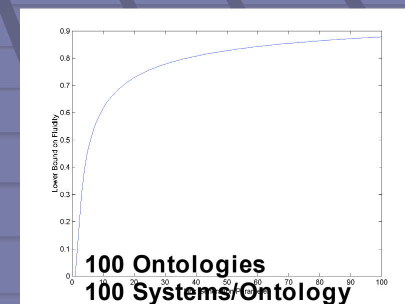
May even be some math proving efficacy

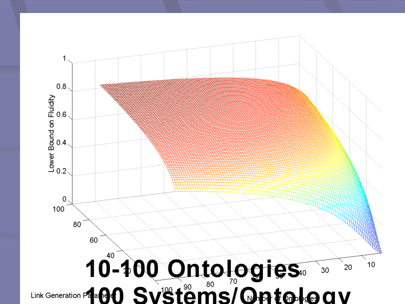




- Some early mathematical results show partial mappings can produce good fluidity with less connections and better overall complexity
  - But *VERY* early results

Jiang, Cybenko, Hendler, 03





MiND

An interdisciplinary vision

... The Semantic Web will provide unifying underlying technologies to allow these concepts to be progressively linked into a universal web of knowledge, and will therefore help to break down the walls erected by lack of communication, and allow researchers to find and understand products from other scientific disciplines. The very notion of a journal of medicine separate from a journal of bioinformatics, separate from the writings of physicists, chemists, psychologists and even kindergarten teachers, will someday become as out of date as the print journal is becoming to our graduate students.

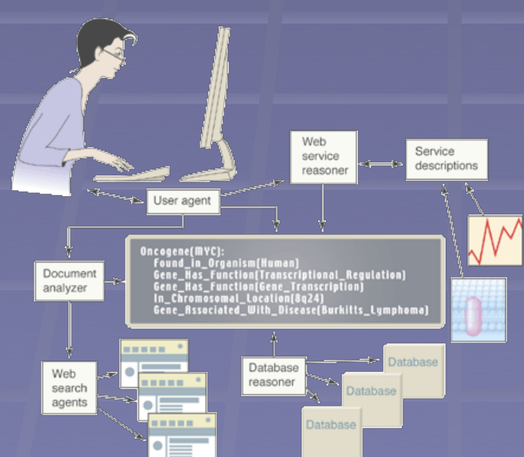
Does this sound like a crazy science-fiction dream? A decade ago, who would have believed a web of text, conveyed by computer, would challenge a 200-year-old tradition of academic publishing?

Publishing on the Semantic Web, Berners-Lee, Hendler, 01

Front End: User tools

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.

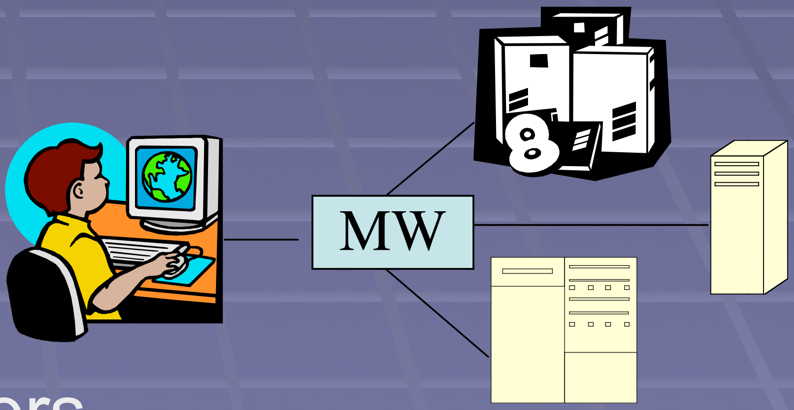
Science and the Semantic Web, Hendler, 03



MiND

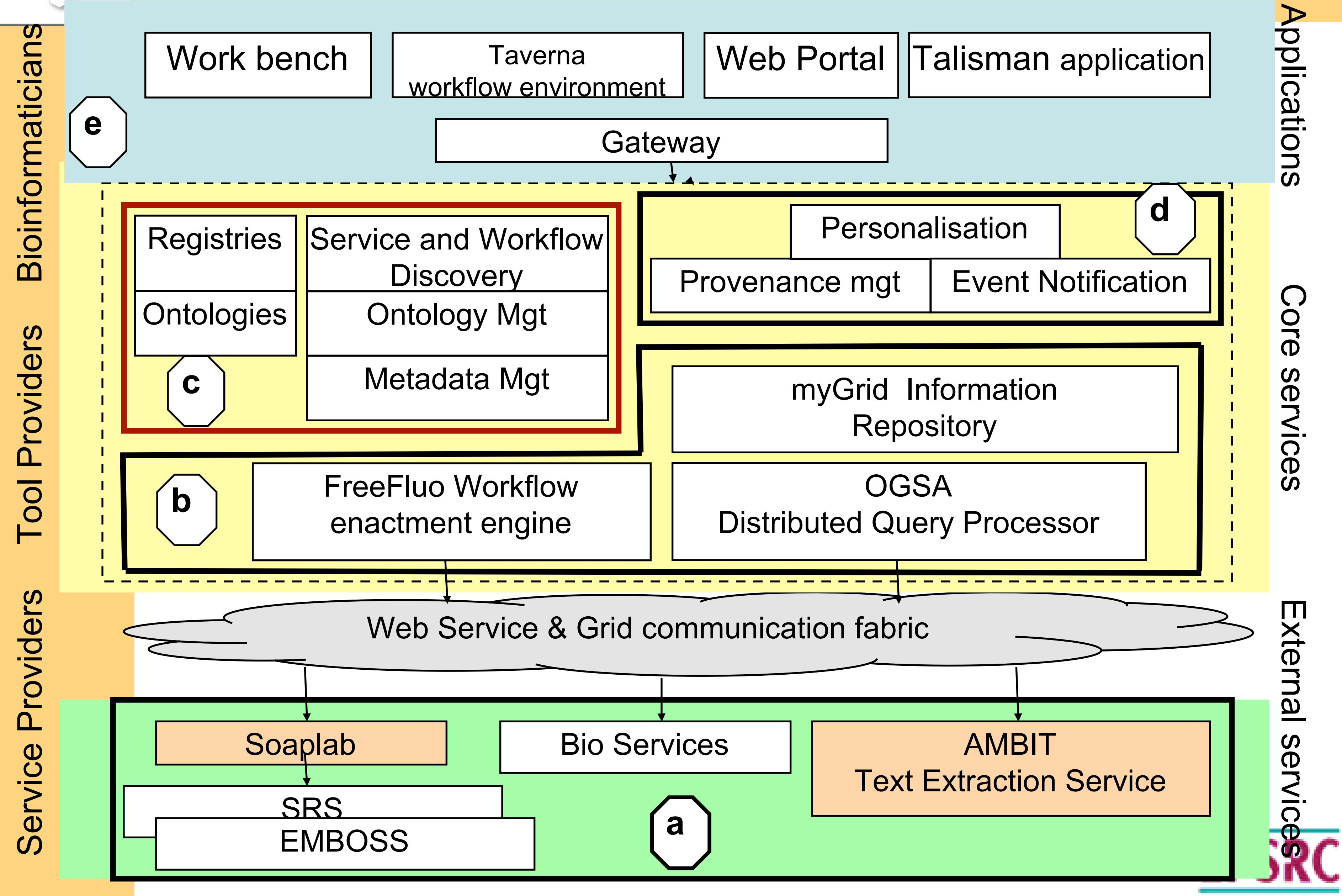
Services

- This slide replaces a whole bunch of slides which talk about middleware, services, and backends, and discuss how applications using service-based middleware provide a mechanism for scientists to access devices, databases, schedulers, etc.
  - OGSA



- Build tools to enable tool building by the scientific instrument/product developers
  - UK E-science researchers already know this!

myGrid Service Stack



OWL-S (nee DAML-S)

- Three core ontologies (Profile, Process Model, and Grounding
  - Supports the discovery (Profile), composition, verification, monitoring (Process Model), and execution (Grounding) of Web Service
- Built on OWL
  - Hard at work on 1.0 release
- OWL-S grounding
  - Likely to be included in WSDL 1.2 recommendation
    - Layered on WSDL 1.1 (though, could work with other IDLs)
    - Describes "how to use" (i.e., the concrete invocation)
    - Point of contact with "the rest" of the Web Service stack
    - Grounding class which contains the mapping

Grounding WSDL

see Luc Moreau's paper in proceedings

input xsd:complex="oncogene"

```xml
- <types>
  - <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="estimatedTotalResultsCount" 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"

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

Lets us add services for even more network effect

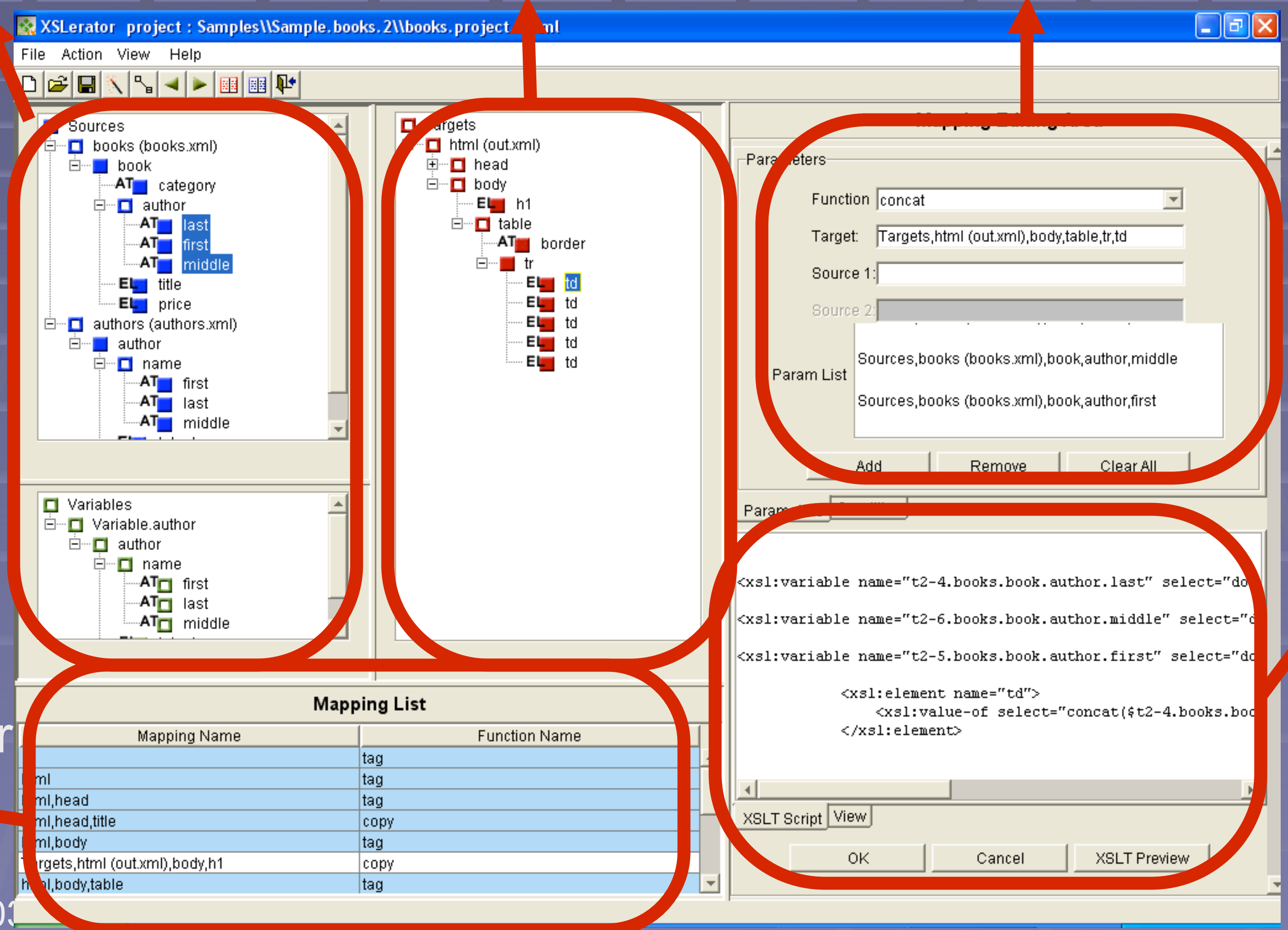
Tools for grounding/linking

DAML Ontology (DAML-S)

XML Schema (WSDL)

Define transformation functions

Joint work with Fujitsu Labs of America



Manual editing if necess.

Mappings for attributes

Semantic Grid Services

COMBINE

Open Grid Services Architecture

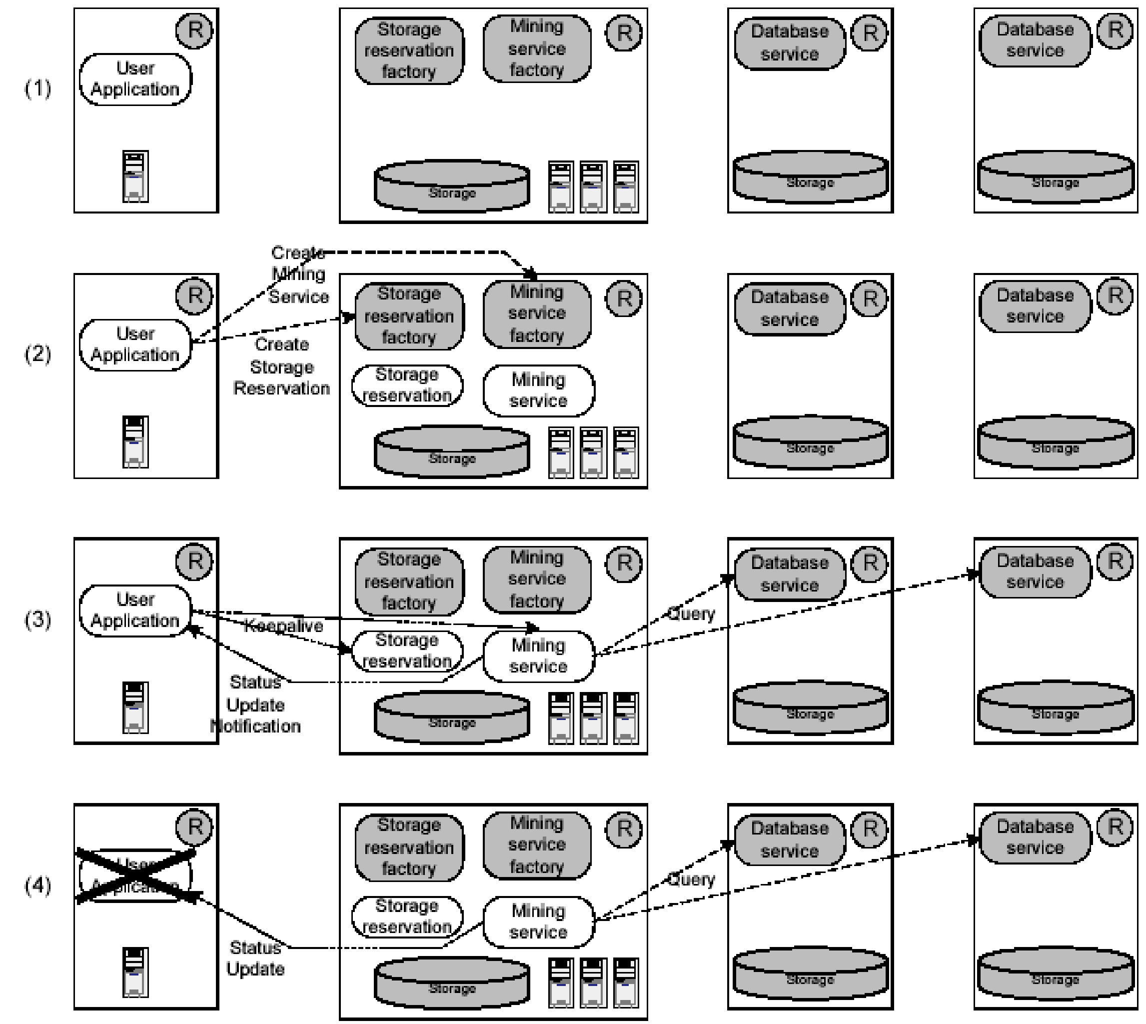


Figure 3: An example of Grid services at work. See text for details.

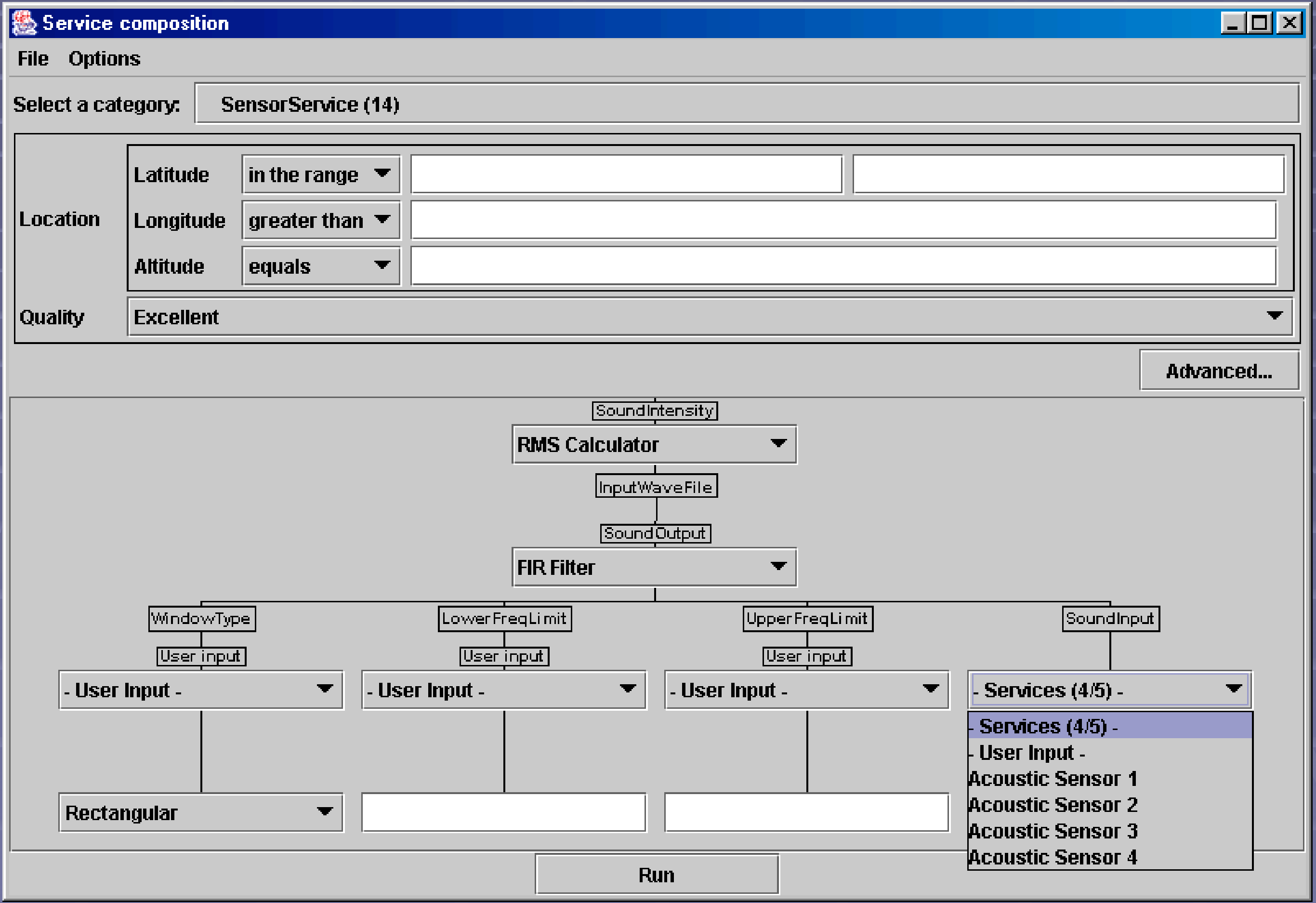
Semantic Grid Services

COMBINE

Open Grid Services Architecture

WITH

Semantic Web Services



Advanced information management capabilities
Discovery, Filtering, Composition

MiND

Service Composition

- OGSA and DAML-S currently play in similar spaces
  - Extending representation of WSDL
  - Grounded in executable WSDL
  - Provide a "formal model" that can be used for
    - Composition
    - Filtering
    - Definition
    - Mapping

UMCP Service Composer (Sirin, Parsia, Hendler, Masuoko, Wu)
- Developed for the Web Service Domain; extended to "sensor/consumer" model for simple example of moving towards OGSA capabilities
- Ontolink developed to demonstrate mapping/service composition dynamics

(Work Joint with Fujitsu Laboratories of College Park)

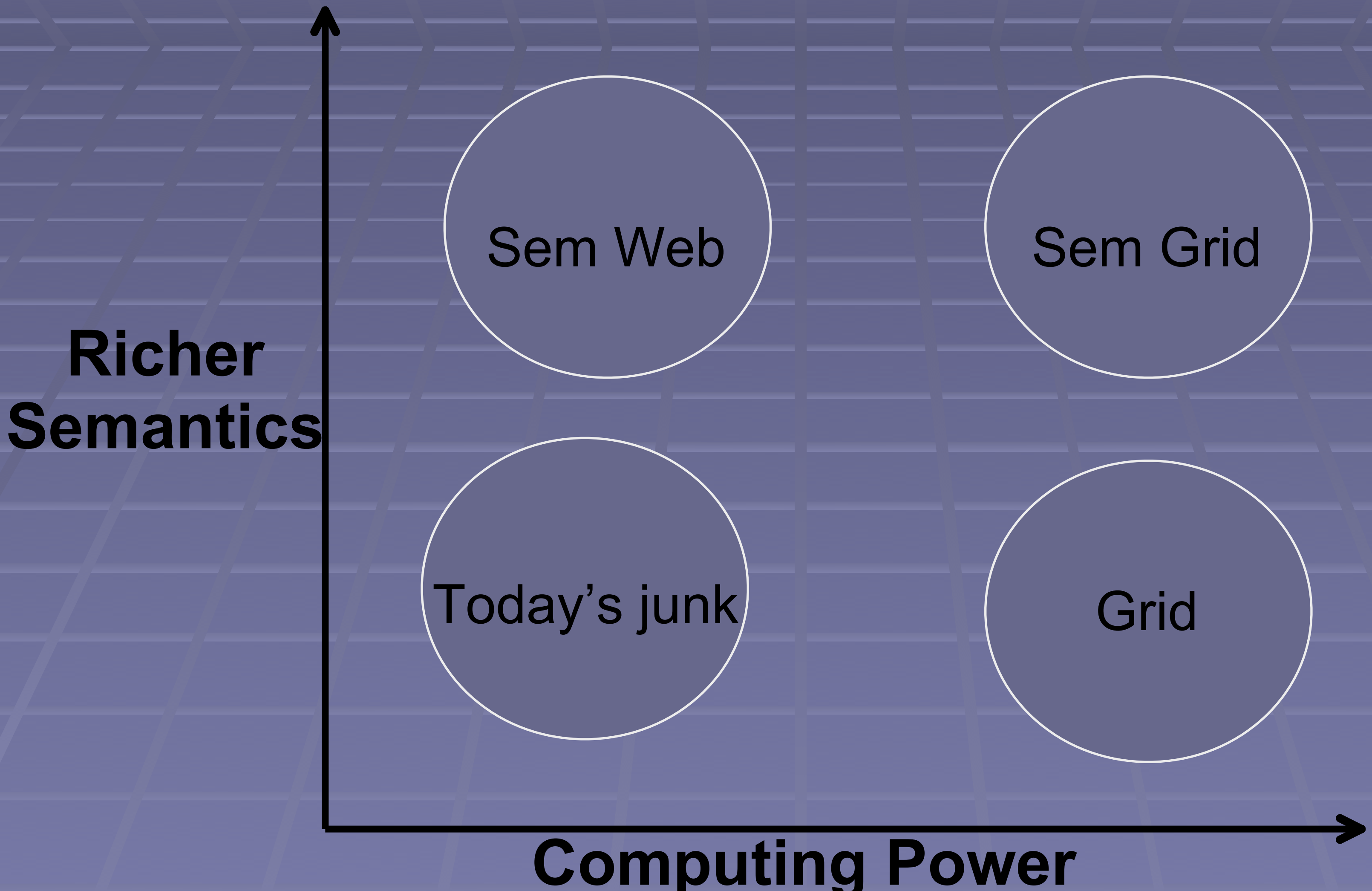
What’s missing?

- Neither OGSA nor DAML-S offers powerful enough process model
  - W3C Web Services Choreography considering a declarative choreography model
    - Pi-calculus possible base
  - OGSA looking at “process/resource” models
    - Essentially extended workflow
  - SWSI considering temporal logic
  - OASIS TC looking at WSBPEL (nee BPEL4WS)
    - Turing complete “programming” model of orchestration
- None of these seem sufficient
  - A key, and mostly ignored, research need
  - Can OGSA/ SWGrid/E-science drive this?
    - important to standardize, but not necessarily to create a formal standard

OK, what about e-science

- E-science clearly a compelling use case for both SemWeb
  - Gene ontology, Galen, NCI Ontology (in OWL!)
  - Medical thesauri
  - Proteomics ontology underway
- And Grid
  - Genomic database analysis/visualization
  - Scientific computation/modeling
- But, exercising them in different ways…

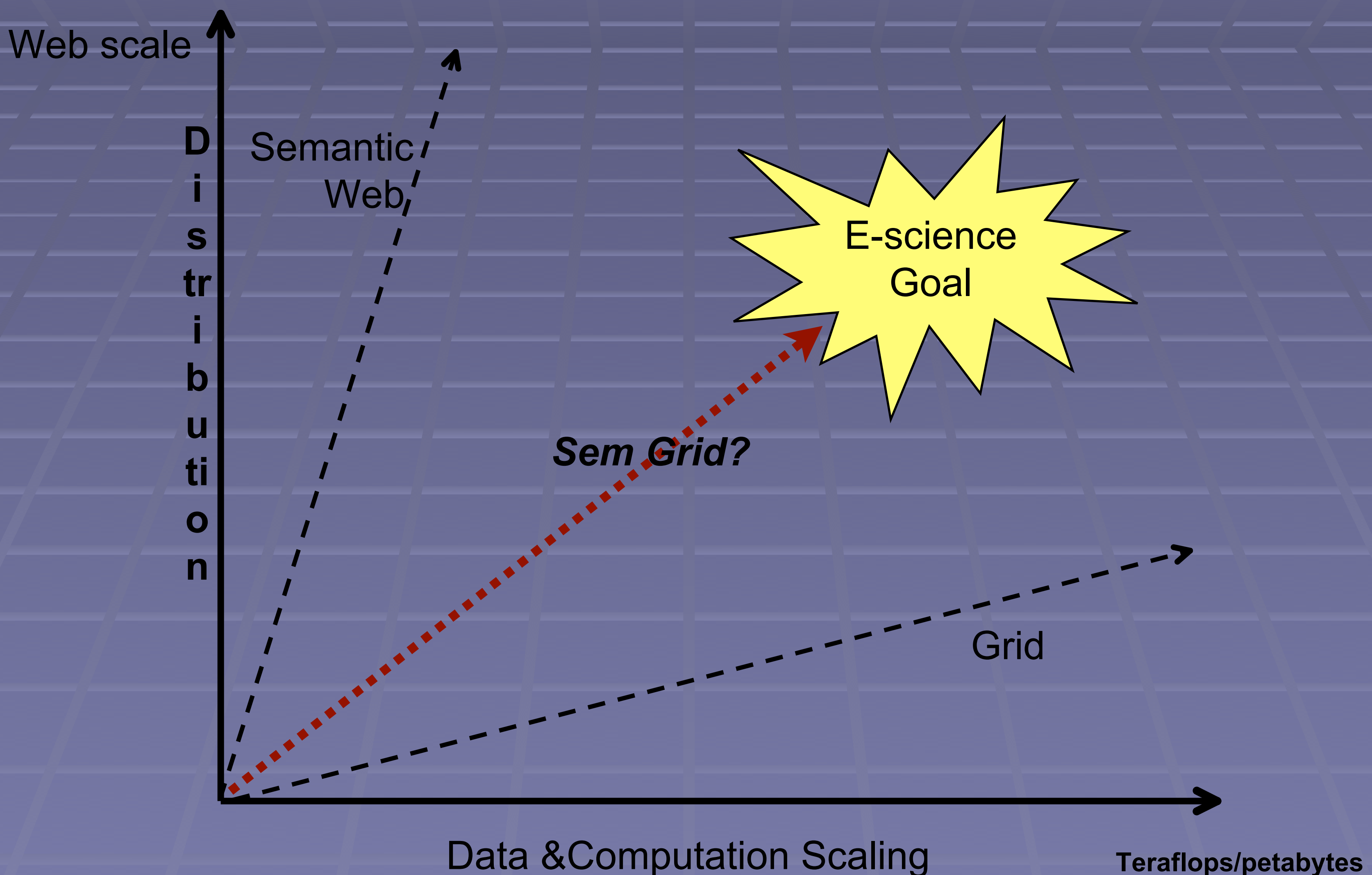
The Semantic Grid ?



Paton, de Roure, Goble, Hey

MiND

The Grid *vs* the Semantic Web



Conclusion

- Semantic Web offers powerful new web technologies for e-science and collaboration
- Grid and Sem Web capabilities bring e-science community to the web
  - Growing emphasis on services and information management -- the Semantic Web's key competencies
- Promising long-term research directions
  - Information **models** on the Web/Grid
  - Integration of Grid/Sem Web services

If we don't integrate Semantic Web ***technologies with*** Grid ***computing, e-science will not succeed***

MIND SWAP

- Maryland Information and Network Dynamics Laboratory, ***Semantic Web and Agents Project***
  - J. Hendler
  - B. Parsia
  - Jennifer Golbeck
  - Aditya Kalyanpur
  - Grecia Lapizco-Encinas
  - Katy Newton
  - Evren Sirin
  - Ronald Alford
  - Ross Baker
  - Amy Alford
  - Matt Westhoff
  - Kendall Clark
  - Nada Hashmi
- Corporate Research Partners:
  - **Fujitsu Laboratory of America, College Park**
  - Lockheed Martin Advanced Technology Laboratories
  - NTT Corp
  - SAIC Corp.
- Govt Funding:
  - US Army Research Laboratory, NSF, DARPA, Agency we mayn't name
- ***http://owl.mindswap.org*** (OWL-powered Semantic Web page)