The Noosphere from the bottom up.

Jim Hendler

RPI

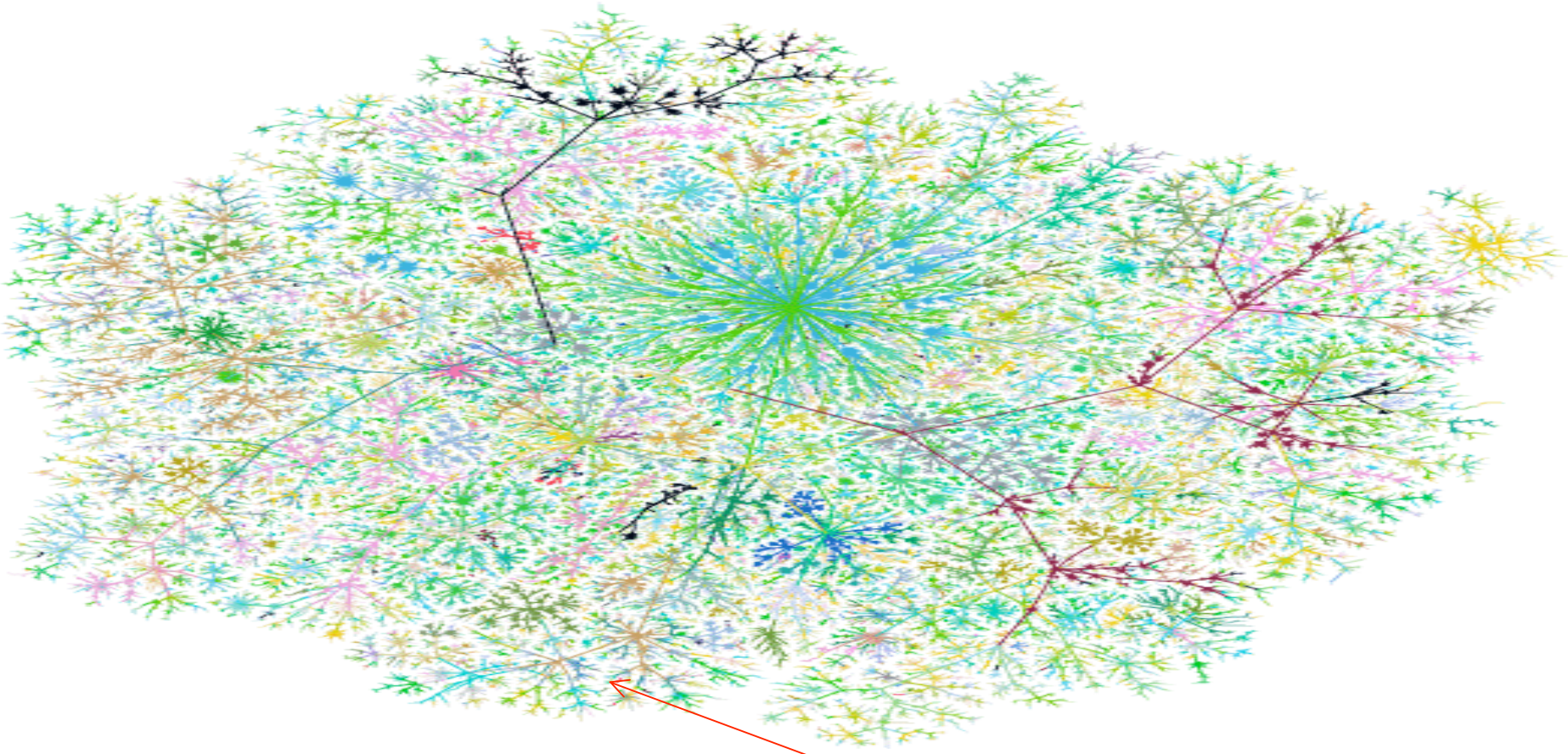
http://www.cs.rpi.edu/~hendler

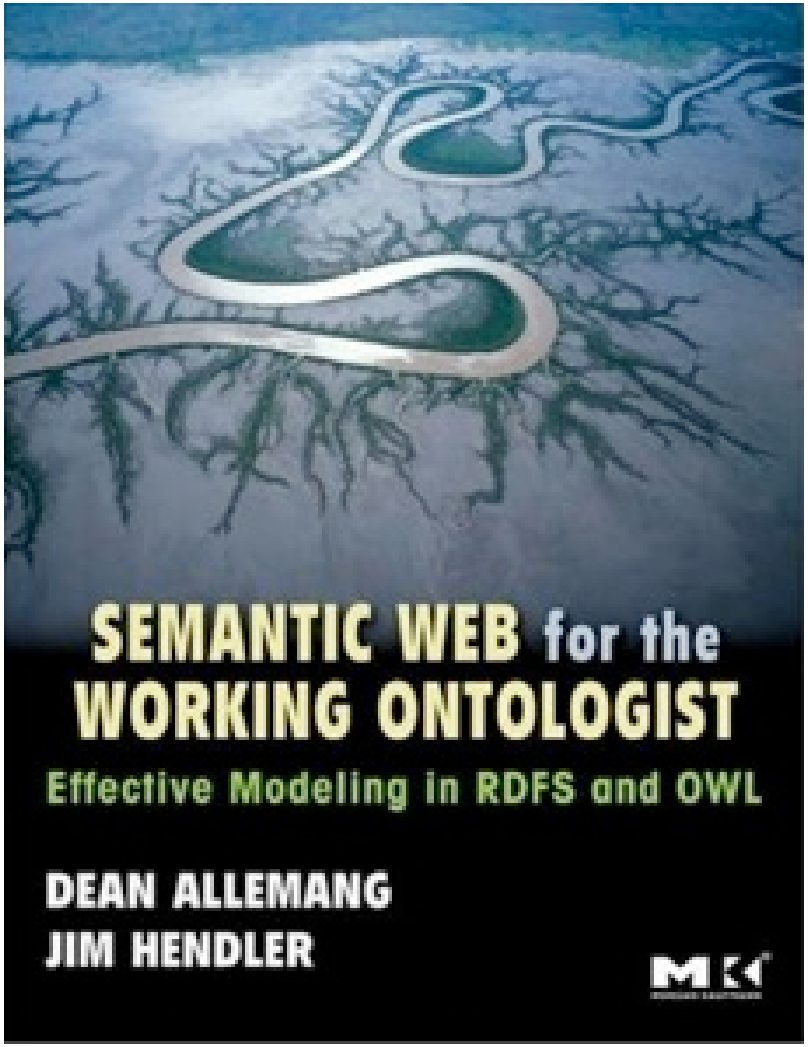
Why am I here?

- The **noosphere** can be seen as the "sphere of human thought" (Wikipedia)
  - (I looked in citizendium, but it wasn't there)
- And:
  - Ontology: provide a definitive and exhaustive classification of entities in all spheres of being" (Smith 2003).
- So I guess that's why I was invited
  - Although I prefer, and will defend, a different view

Ontology: an abstract, simplified view of the world that we wish to represent for some purpose (Gruber 1995).

The Web is Humongous





http://www.amazon.com/Semantic-Web-Working-Ontologist-Effective/dp/0123735564

Rensselaer

Tetherless World

Working ontologist?

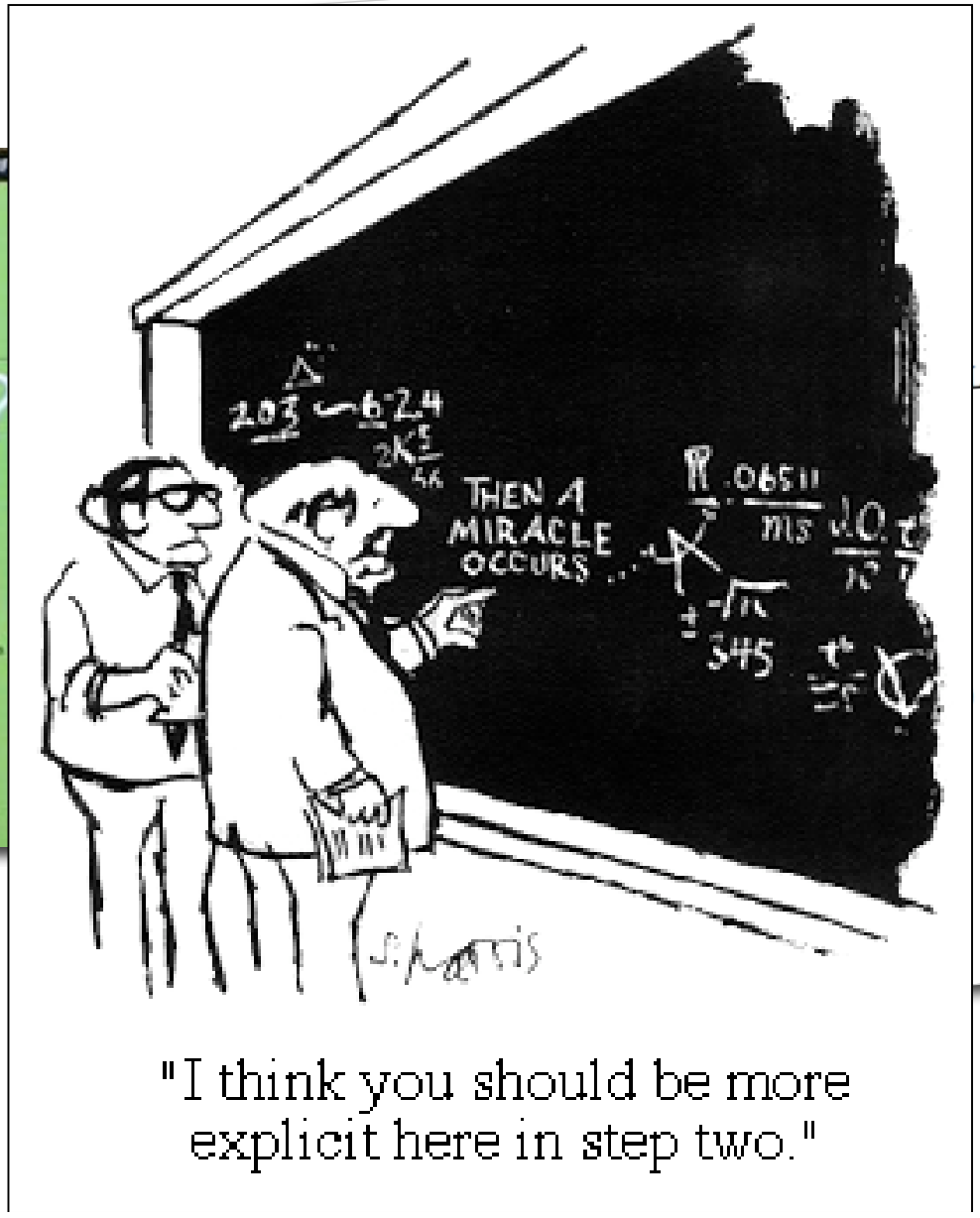
"The solution to any problem in AI may be found in the writings of Wittgenstein, though the details of the implementation are sometimes rather sketchy." (Hirst, 2000)

- This talk aims at exploring "implementation details" for "knowledge" on the Web
  - Individual and Collective

Knowledge

- Philosopher: what is it?
- Cognitive Scientists: how do humans learn/use it?
- Social Scientist: how does it manifest in behavior
- Computer Scientist: What can I do with it?
  - Web Scientist: What can I do with it on the Web?

The Semantic Web (ca. 2001)



Rensselaer

Tetherless World

Semantic Web ca. 2008

- **Semantic Web** companies starting & growing
  - Siderean, SandPiper, SiberLogic, Ontology Works, Intellidimension, Intellisophic, TopQuadrant, Data Grid, Mondeca, ontoPrise…
  - Web 3.0 new buzzword: Garlik, Metaweb, RadarNetworks, Joost, Talis, …
- Bigger players buying in
  - Adobe, Cisco, HP, IBM, Microsoft, Nokia, Oracle, Sun, Vodaphone, Yahoo!, Reuters, …
  - Gartner identifies Corporate Semantic Web as one of three "High impact" Web technologies
  - Tool market forming: AllegroGraph, Altova, TopBraid, …
- Government projects in and across agencies
  - US, UK, EU, Japan, Korea, China, India…
- Several "verticals" heavily using Semantic Web technologies
  - Health Care and Life Sciences
    - Interest Group at W3C
  - Financial services
  - Human Resources
  - Sciences other than Life Science
    - Virtual observatory, Geo ontology, …
- Many open source tools available
  - Kowari, RDFLib, Jena, Sesame, Protégé, SWOOP, Pellet, Onto(xxx), Wilbur, …

SW now becoming "visible" on the Web

Rensselaer

Tetherless World

Why not just words, statistics and web pages

- While it is true that the relations between words and/or the contexts they appear in can be powerful
  - *cf.* Links and context power Google
  - *cf.* Wordnet (more precision in definition)
  - *cf.* Powerset *(now part of Microsoft)*
- That only goes so far
  - *Ambiguity*
  - *Symbol Grounding*
  - *Personalization and individual differences*
  - *Non-linguistic resources (images, video, data)*

Rensselaer

Tetherless World

Traditional AI Knowledge Representation

- Relation between contents can be defined as logical entailments in a formal system
  - Student(?x) => Person(?x)
  - In this view, *ontology* is defined as the formal domain model for some segment of the world
- Which is often criticized (rightly) for
  - Complexity/Undecidability
  - Definitional adequacy
  - Knowledge Engineering bottleneck
  - Grounding

Seeing new life as part of the Semantic Web

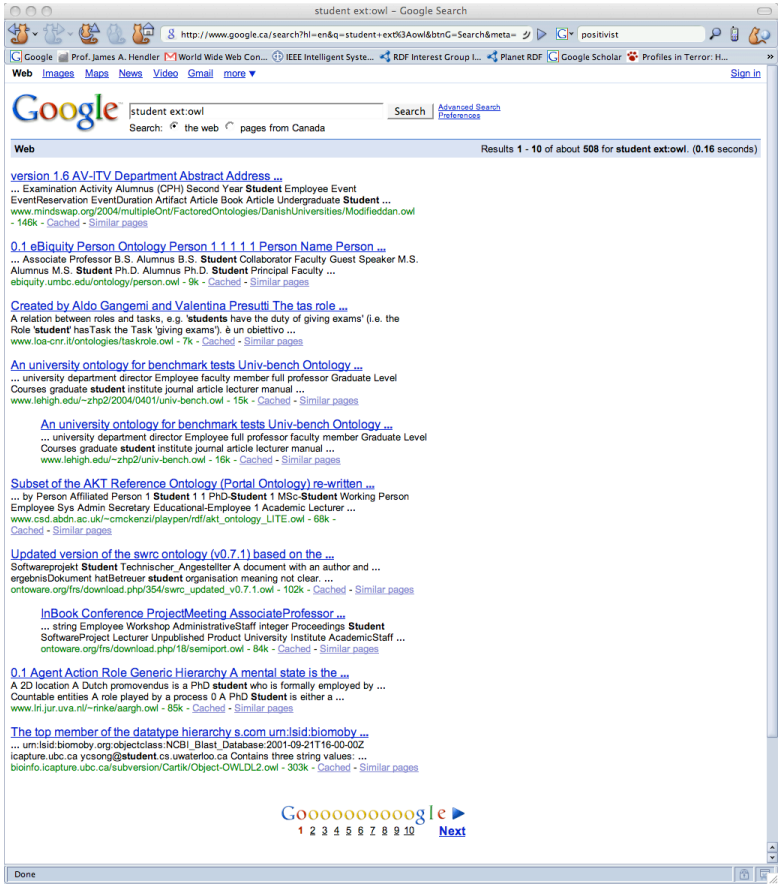
- Web ontology language OWL
  - A small set of terms, formally defined to produce specific entailments
    - i.e. given some facts, specify the *mandated* entailments (All and Only)
    - A standard for the Web
      - High buy in from many in the "KR" community
      - Some buy in from many in the Web Application community
    - Most used KR language in history (by many orders of magnitude)
      - Depending on how you ask, Google finds thousands to tens of thousands OWL ontologies

So why is it working this time?

Rensselaer

Tetherless World

Google for "student ext:owl"



...

Rensselaer

Tetherless World

Widely Varying Quality

- *cf.* US National Center for Biotechnology Information, "Oncology Metathesaurus"
  - 50,000+ classes, ~8 people supporting full time, monthly updates, mandated for use by NIH-funded cancer researchers
    - OWL DL rigorously followed
    - Provably consistent
- *cf.* Friend of a Friend (Foaf)
  - 30+ classes, Dan Brickley and Libby Miller made it, maintained by consensus in a small community of developers
    - Violates DL rules (undecidable)
    - Used inconsistently

Widely varying use

- NCBI Oncology Ontology
  - High use in medical community
  - High cost for specific representational need
  - Not much data on the web
- FOAF
  - ~60M Foaf people (not necessarily distinct individuals)
  - Used by a number of large providers
    - If you use LiveJournal, you have a FOAF file
      - Also flickr, ecademy, tribe, joost, …
      - And you can export Foaf from Facebook and many other social networking sites
  - Becoming de facto standard for open social networking

Why?

- CLAIM SET 1: Formal properties
  - Based on a decidable subset of KR
    - Description logics
  - For which much scaling research has been happening
    - *Ca. 2000 - 10,000 axioms, no facts, 1 day*
    - *Ca. 2008 - 50,000 axioms, million facts, 10 min.*
      - Not just faster computers (but Moore's Law helps), significant research into optimization, "average case"
      - Moving to parallel (Web server)
  - With some new ways of linking to larger data sets
    - SHER, IBM, "reduced Abox"
    - OWL-Prime, Oracle, "materialized views"

In this view OWL is a formal KR standard

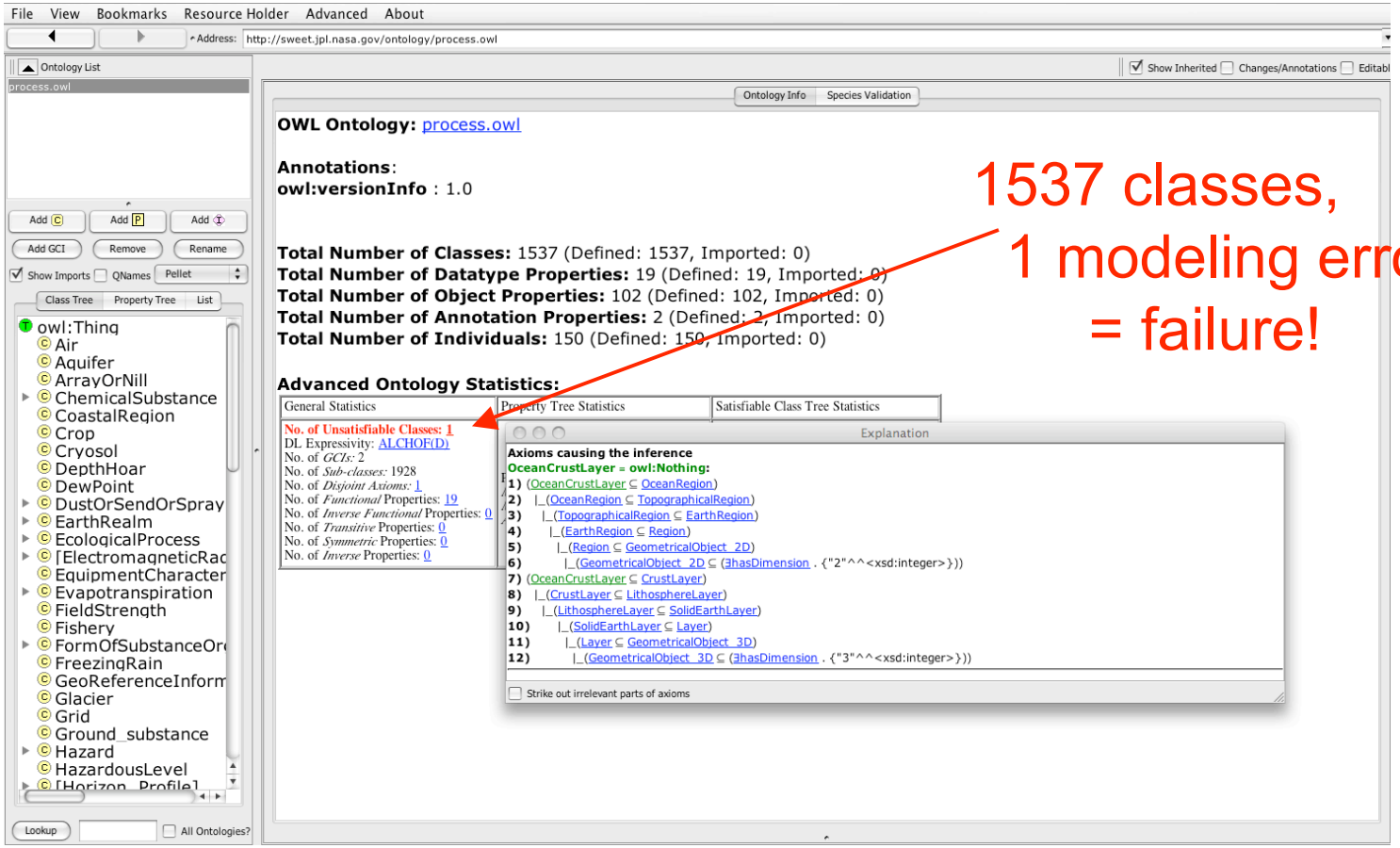
O*ntology*: the formal KR view

- **O**ntology as Barad-Dur (Sauron's tower):
  - Extremely powerful!
  - ~~Patrolled by Orcs~~ Decidable Logic basis
    - Let one little ~~hobbit~~ inconsistency in, and the whole thing could come crashing down

Rensselaer

Tetherless World

Inconsistency is the bane of this view



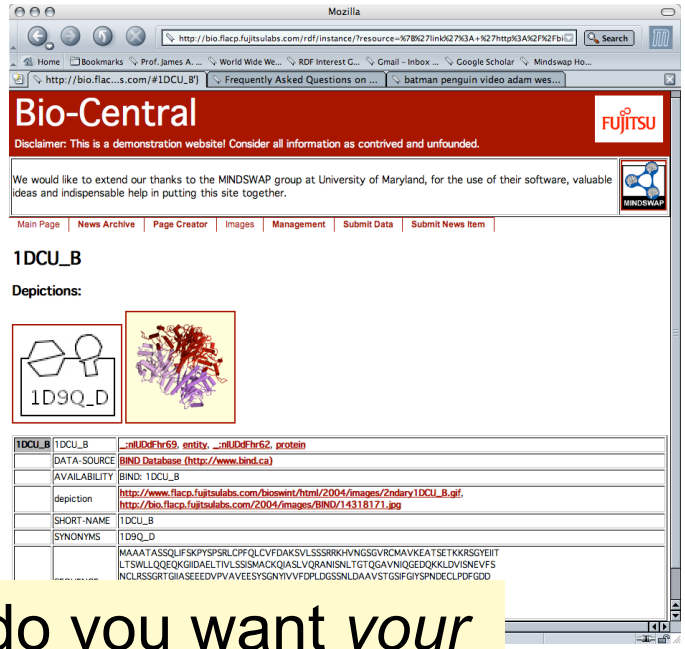
(Swoop w/Pellet)

Rensselaer

Tetherless World

The argument for this is often compelling

- When "folksonomy" isn't enough…



Which one do you want *your* doctor to use?

Rensselaer

Tetherless World

Goal: Reasoning over (Enterprise) data

- Formal modeling finds its use cases in verticals and enterprises
  - Where the vocabulary can be controlled
  - Where finding things in the data is important
- Example
  - Drug discovery from data
    - **Model** the molecule (site, chemical properties, etc) as **faithfully** and expressively as possible
    - Use "Realization" to categorize data assets against the ontology
      - **Bad or missed answers are money down the drain**
- But the modeling is very expensive and the return on investment must be very high!
  - Which is part of why the "expert systems revolution" wasn't one
  - Became part of the technology tool kit, a useful niche in the programming pantheon, but didn't change the world

The alternative

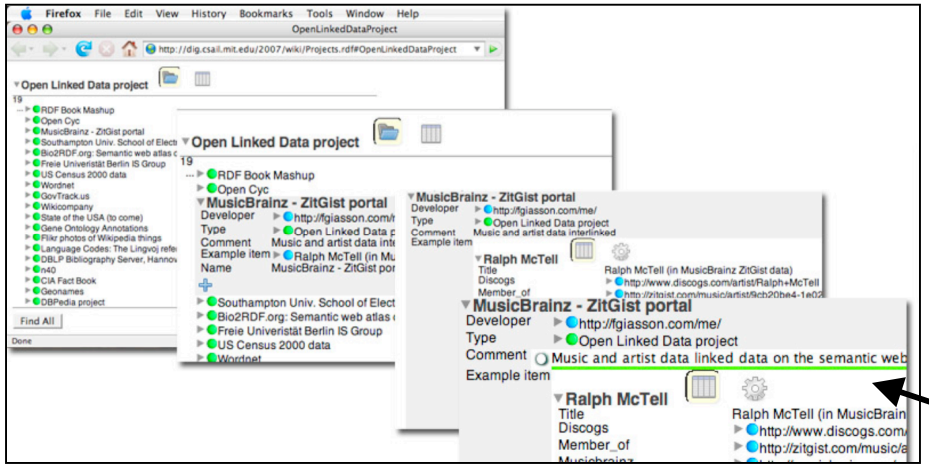
- OWL is based on RDF, a language designed for the (Semantic) Web
  - Built with Web architecture in mind
    - Exploits Web infrastructure, respects W3C TAG recommendations
      - Internationalization, accessibility, extensibility
  - Fits the ***Web culture***
    - Open and extensible, supports communities of interest
      - *If you don't like my ontology, extend it, change it, or build your own*
    - Fits the Web application development paradigm
      - Scales like "databases"
  - With some new ways of linking to formal models
    - Heavy use of a small amount of OWL
    - Generally used "like it sounds" not like the formal model
      - Example "owl:sameAs" debate

OWL is a "webized" ontology language

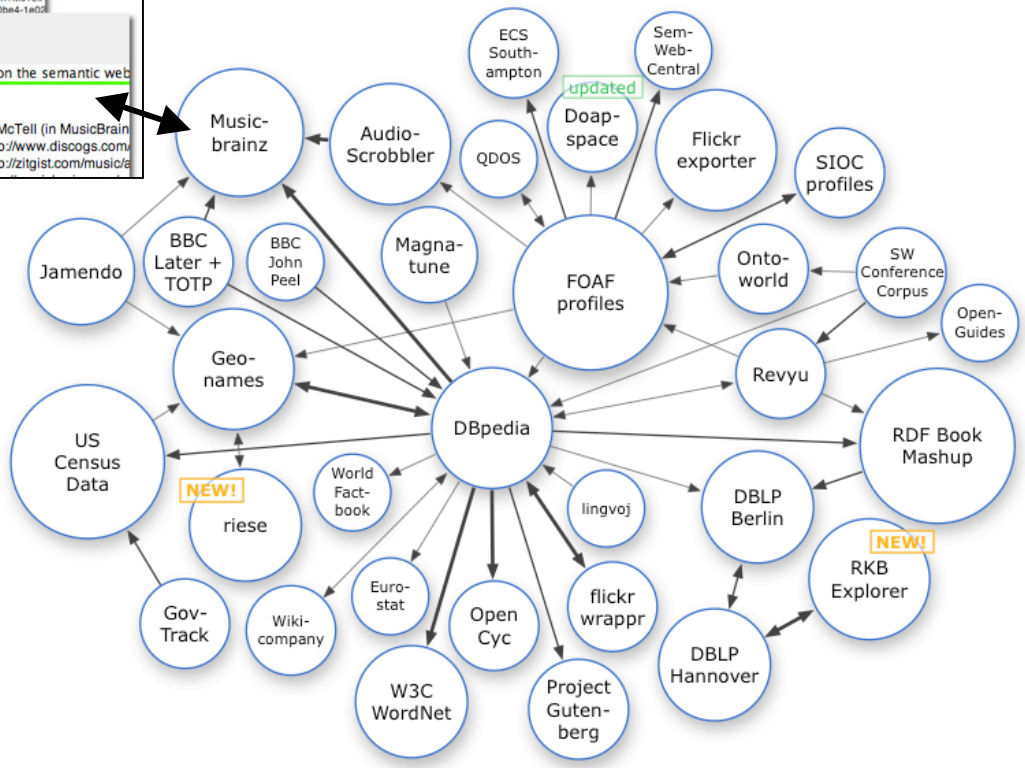
Rensselaer

Tetherless World

Navigate between ontologies/datasets without boundaries



Tabulator and Linked Open Data



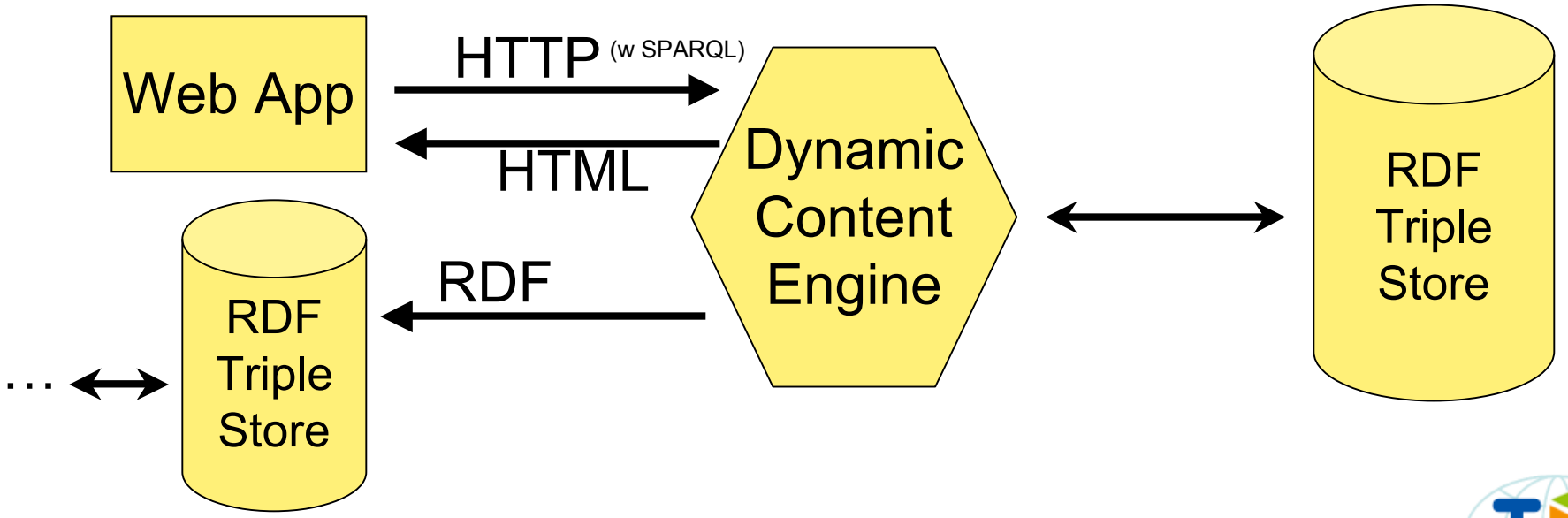
Rensselaer

Goal: create "Web 3.0"

- "Data Web" approach finds its use cases in Web Applications (at Web scales)
  - A lot of data, a little semantics
  - Finding anything in the mess can be a win!
- Example
  - Declare simple inferable relationships and apply, at scale, to large, heterogeneous data collections
    - *eg.* Use InverseFunctional triangulation to find the entities that can be inferred to be the same
      - These are "heuristics" not every answer must be right (qua Google)
      - But remember *time* = *money*!

Web 3.0 is happening

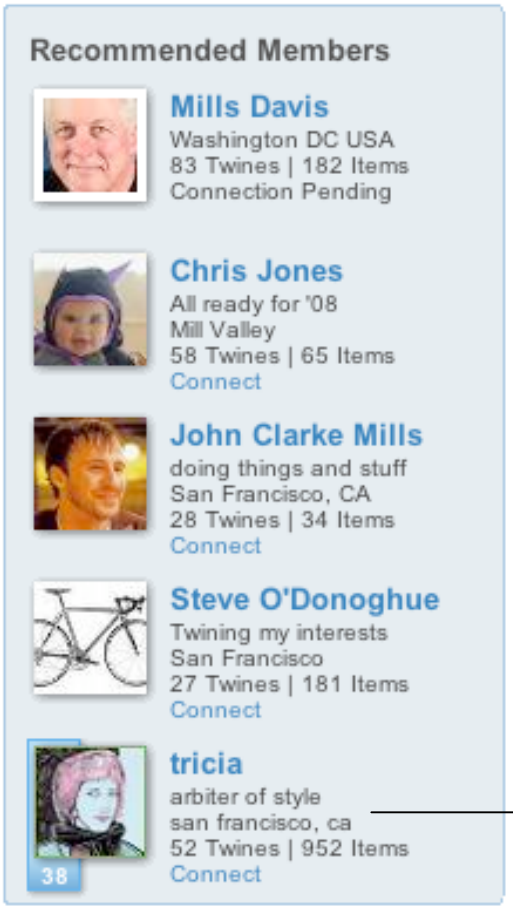
- ~2006: Web app developers discover the Semantic Web



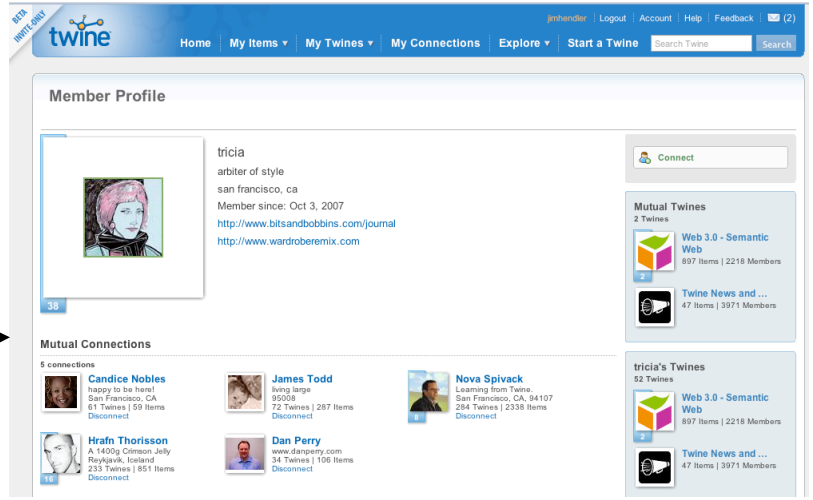
Rensselaer

Tetherless World

How do these applications ignore completeness?



- Twine recommends some people I may want to connect to
  - What is correctness in this case?
    - If I find some folks I like this way, I use twine more. Surprises can be fun.
  - I'm only seeing a few of a very large set (think Google) so "first" is more important than "there somewhere"



Rensselaer

Tetherless World

o*ntology*: the Webbie view

Genesis 11:7 Let us go down, and there confound their language, that they may not understand one another's speech. So the Lord scattered them abroad from thence upon the face of all the earth: and they left off to build the city.

- …logy and the …f Babel
  - …build a tower …ach the sky
  - …e only need a little ontological agreement
    - Use Wordnet or other "linguistic" constructs

Rensselaer

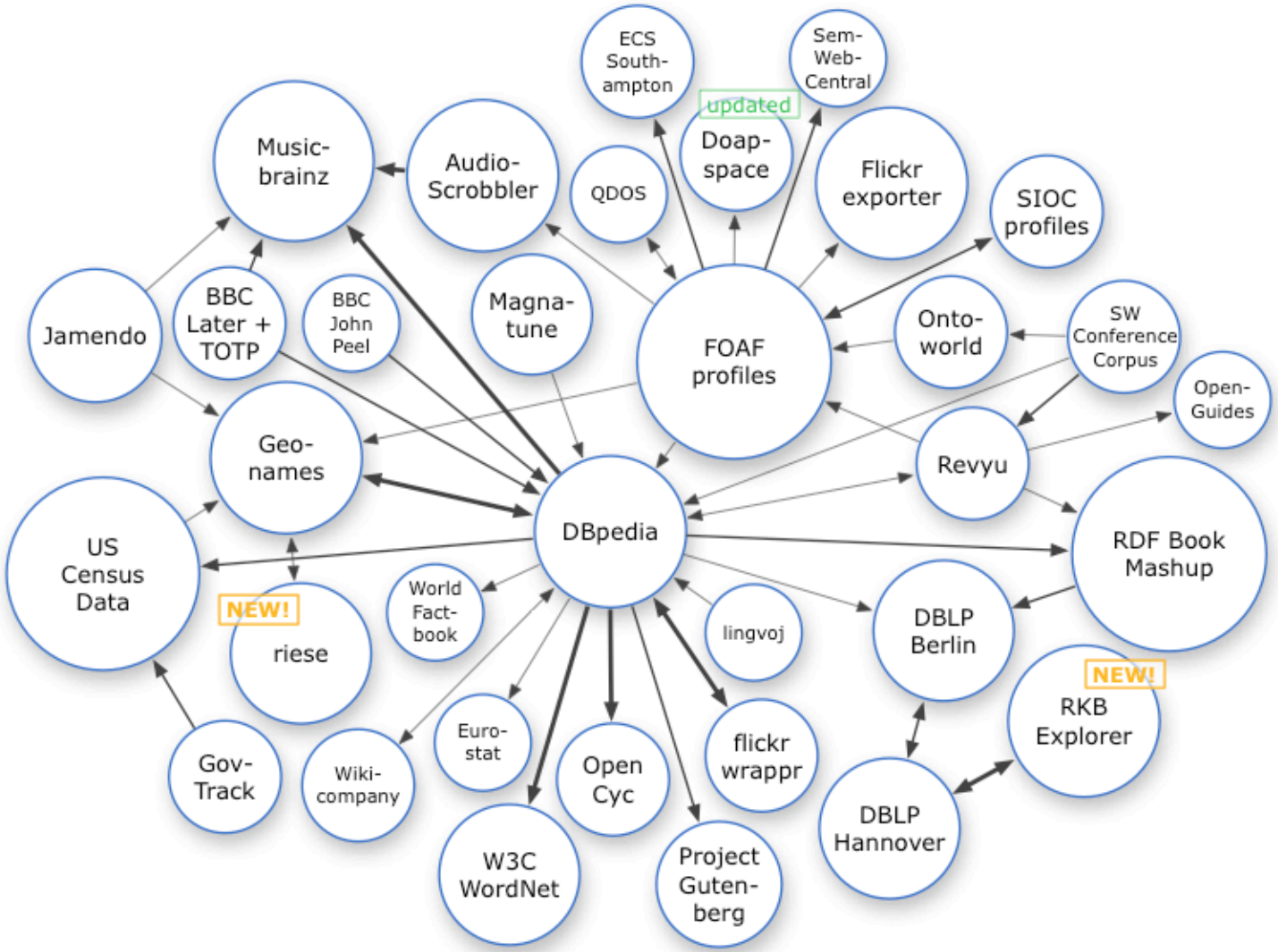
Tetherless World

Avoiding Babel

- The essential process in *webizing* is to take a system which is designed as a closed world, and then ask what happens when it is considered as part of an open world. Practically, this effect on a computer language is to replace the names/ tokens/identifiers for URIs. Thus, where before reference could only be made to something in the same document/ program/module one can with equal ease make reference to something in a different one somewhere in that abstract space which is the Web. (Berners-Lee, 1998)

Advantages

- Why ground terms in URIs?
  - "student" ≠ http://www.cs.rpi.edi/~hendler/Twgroup.owl#student
    - A talk in itself (or a debate with Stevan Harnad)
  - Can recognize equality (same URI = same concept)
  - Can assert equality (URI1 owl:sameAs URI2)
  - Can assert inequality (URI1 owl:differentFrom URI2)
  - Can combine (URI1 foaf:depicts URI2 foaf:name "Jim Hendler")
- Other advantages
  - Infinitely extensible name space
  - Can be dereferenced
    - Click on the term, see the definition (and thus know the entailments)
    - Ubiquitously implemented (from server to phone)
  - Well understood ***social conventions***
    - RPI's server maintains, and user hendler controls, the URI above
  - And can be displayed in any browser anywhere in the world
    - (and w/labels in different languages, character ses, etc.)



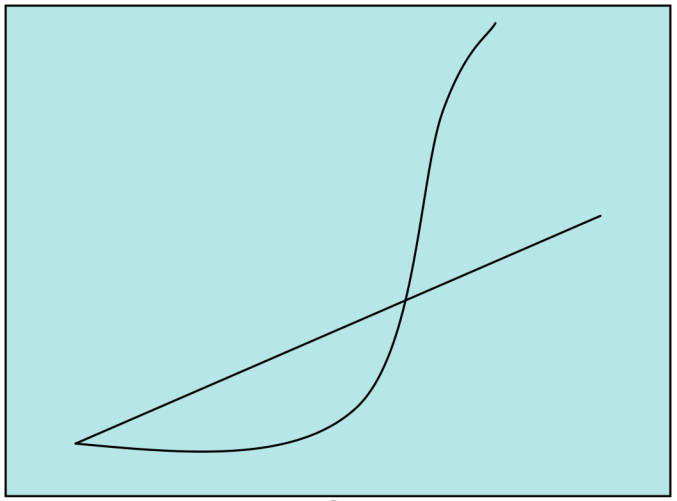
The linked open data cloud now has billions of assertions, and is growing rapidly

Rensselaer

Tetherless World

Linking is power

- Today we can find thousands of ontologies
  - Available on the Web
    - Linked to Web resources
    - Linked to data resources
    - Linked to each other
    - Linked to Web 2.0-like annotations
- And billions of annotated (semi-Knowledge engineered) objects
  - Available on the Web
    - Linked to Web resources
    - Linked to data resources
    - Linked to each other
    - Linked to the ontologies
- Many Large (and curated) "Vocabularies" for Grounding Applications
  - Natl Library of Agriculture (SKOS)
  - NCI Ontology (OWL)
  - Getty Catalog (OWL, licensed), UMLS (RDFS, licensed),
  - GeoNames (RDF), PlaceNames (OWL, proprietary)
  - …

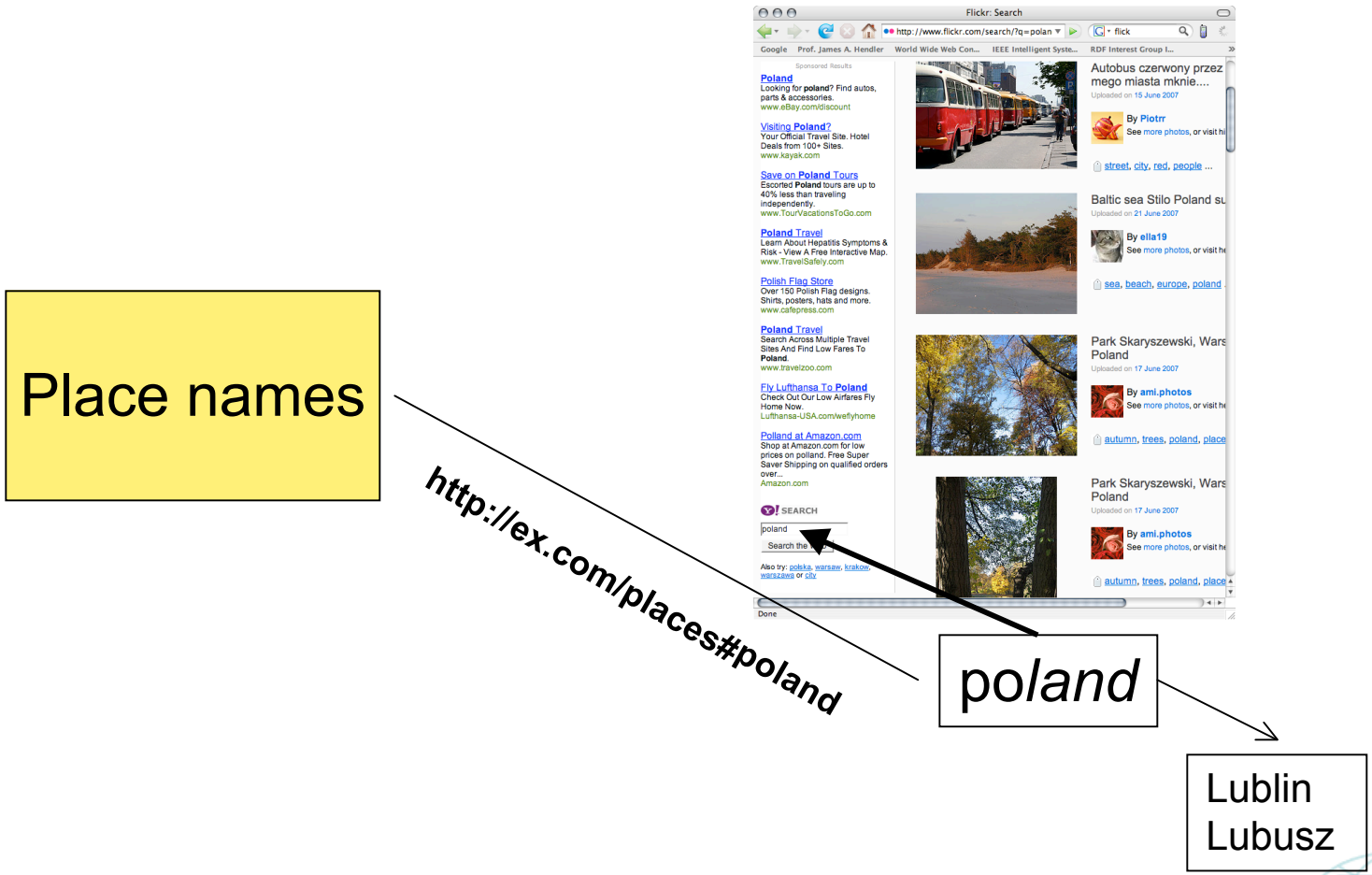


Metcalfe's Law

Rensselaer

Tetherless World

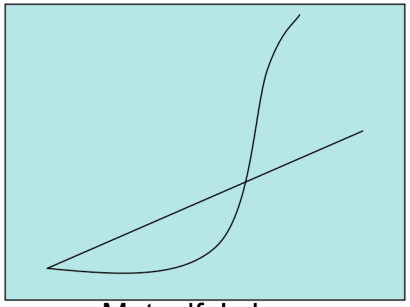
Example: Seeded tagging



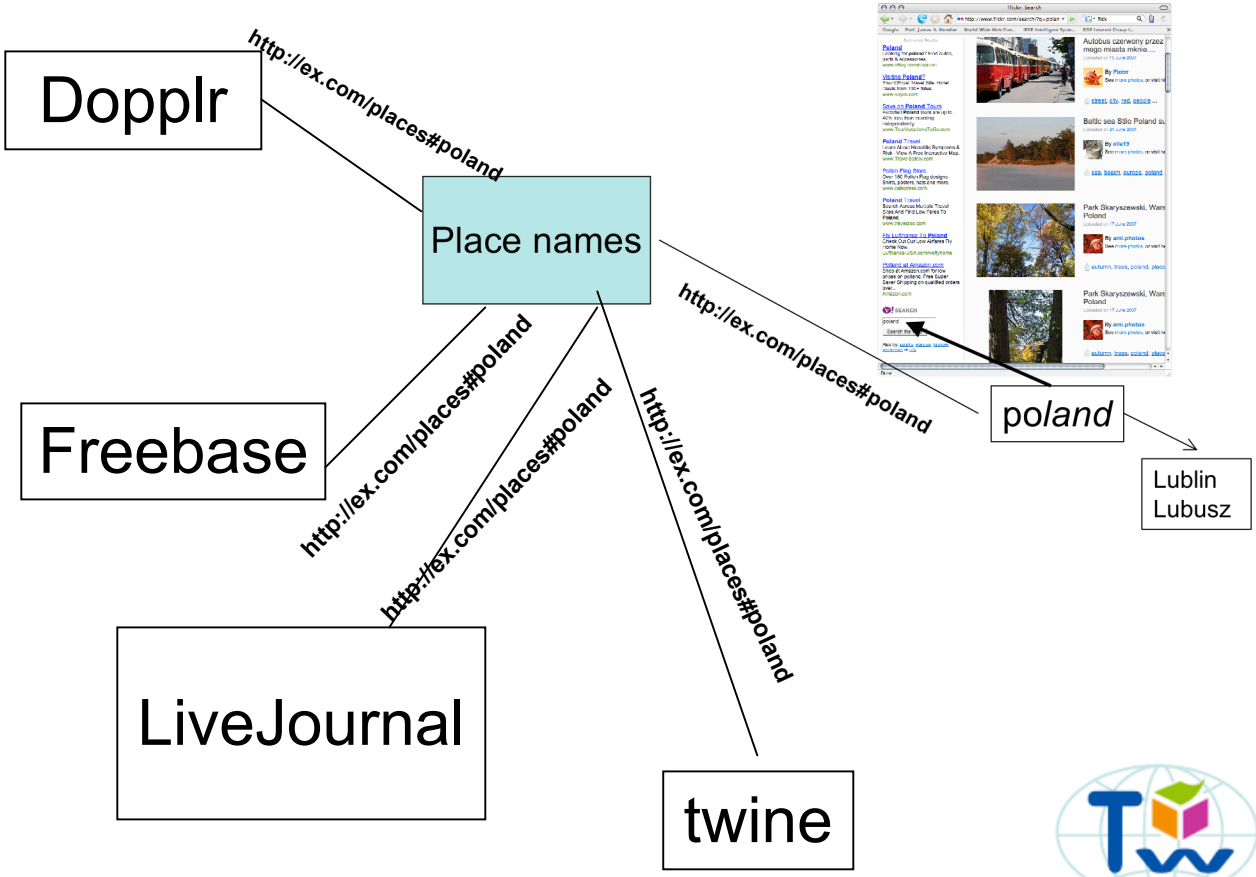
Rensselaer

Tetherless World

Network Effect



Metcalfe's Law



Rensselaer

Tetherless World

The wine ontology (wine.owl)

- Original view: Consensus knowledge of wine and food
  - Lots of debate in its creation
  - Eventually completed with "correct" wine recommendations
    - You disagree, tough! You're wrong.

Wine Ontology Take II

Rensselaer

Tetherless World

TW Wine Agent

Overview
Acknowledgements

Why MountEdenVineyardEdnaValleyChardonnay was selected for Fish

Wine Properties

NAME: MountEdenVineyardEdnaValleyChardonnay
COLOR: White
BODY: Medium
FLAVOR: Moderate
SUGAR: Dry

List of recs being considered

Supporting Recs

TOTAL IN SUPPORT: 9

ID	COLOR	BODY	FLAVOR	SUGAR
MountEdenVineyardEdnaValleyChardonnay	White	Medium	Moderate	Dry
Bland-2Dfish	White	Medium U Full	Moderate U Strong	
RecDLM Swordfish	White	Medium	Moderate	Dry
RecDLM Tuna	White	Medium	Moderate	Dry
RecSwordfish	White	Medium		
RecNonBlandFish	White		Moderate	
RecDLM NonBlandFish	White	Medium	Moderate	Dry
RecFish	White		Moderate	Dry
RecDLM Fish		Medium		Dry
RecSeafood	White			

Opposing Recs

TOTAL IN CONFLICT: 6

ID	COLOR	BODY	FLAVOR	SUGAR
MountEdenVineyardEdnaValleyChardonnay	White	Medium	Moderate	Dry
RecDLM_Scrod	White	Medium	Delicate ⊗	Dry
Melville_Estate_Chardonnay_2006	White	Light ⊗	Strong ⊗	Sweet ⊗
RecDLM_Halibut	White	Medium	Delicate ⊗	Dry
Rec-2Dhendler	Red ⊗	Light ⊗		Dry
RecDLM_Flounder	White	Medium	Delicate ⊗	Dry
RecDLM_BlandFish	White	Medium	Delicate ⊗	Dry

Rensselaer

Tetherless World

TW Wine Agent

Overview
Acknowledgements

Why LongridgeMerlot was selected for Swordfish

Wine Properties

NAME: LongridgeMerlot
COLOR: Red
BODY: Light
FLAVOR: Moderate
SUGAR: Dry

List of recs being considered

Supporting Recs

TOTAL IN SUPPORT: 1

ID	COLOR	BODY	FLAVOR	SUGAR
LongridgeMerlot	Red	Light	Moderate	Dry
Rec-2Dhendler	Red	Light		Dry

Opposing Recs

TOTAL IN CONFLICT: 6

ID	COLOR	BODY	FLAVOR	SUGAR
LongridgeMerlot	Red	Light	Moderate	Dry
RecSwordfish	White ⊗	Medium ⊗		
RecNonBlandFish	White ⊗		Moderate	
Melville_Estate_Chardonnay_2006	White ⊗	Light	Strong ⊗	Sweet ⊗
RecFish	White ⊗		Moderate	Dry
RecDLM_Swordfish	White ⊗	Medium ⊗	Moderate	Dry
RecSeafood	White ⊗			

back

Rensselaer

Tetherless World

The new challenge…

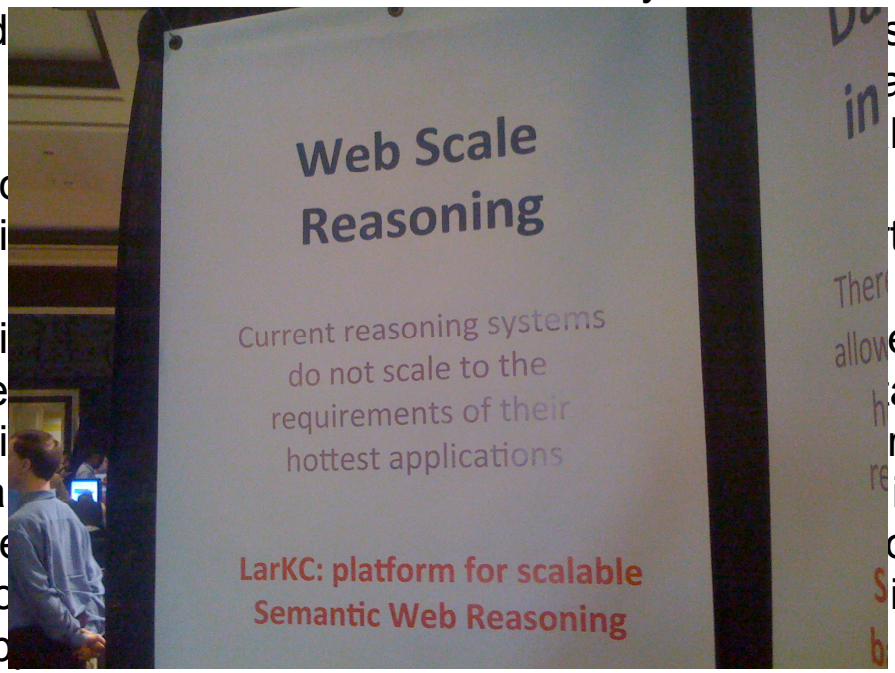
- What do we do with all this stuff?

* The primary goal is to for submissions to show how they add value to the very large triple store. This can involved s in the store via browsing, visualization, ation not directly queriable in the original l information could be tied to part(s) or the wh

* The tool or applicati tion of the data provided by the organizers.

* The tool or applicati ed to the target dataset, but there is still an expe a provided.

* The tool or applicati r application, as defined for the Open Track Cha is to demonstrate an interaction with the large owever, given the scale of this challenge, soluti ications, or as crucial to the success of future ap



(ISWC 2008 - Open Web, Billion Triple Challenge)

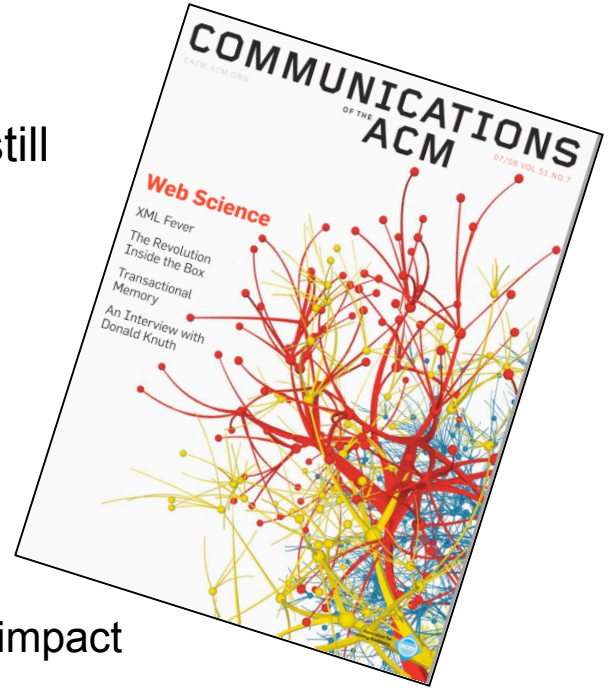
http://iswc2008.semanticweb.org/calls/call-for-semantic-web-challenge-and-billion-triples-tracks/

Rensselaer

Tetherless World

Web Science…

- The Web is a complex and messy place
  - Some "order" added by Semantic Web, but still many avenues of evolution and/or design
- The Web is evolving in many complex ways
  - Today's example, social issues in Web use
    - New functionalities
    - But potentially disruptive technologies
- This drives us towards a new agenda
  - Understanding the Web in a Scientific Way
    - Modeling, engineering, and especially, social impact



- "Web Science", CACM July 2008
  (Hendler, Hall, Shadbolt, Berners-Lee, Weitzner)

Rensselaer

Tetherless World

Summary

- Can use logic, beyond words, on the Web
  - Grounds in URIs
    - Critical! Without it no linking, no network effect
  - RDF/OWL being used on tens of millions of web pages
  - But both formal and "informal" models seem to be emerging
    - New efforts to explore how to link these
  - Semantic Web "visibility," in Web terms, is just beginning
    - Watch this space for more