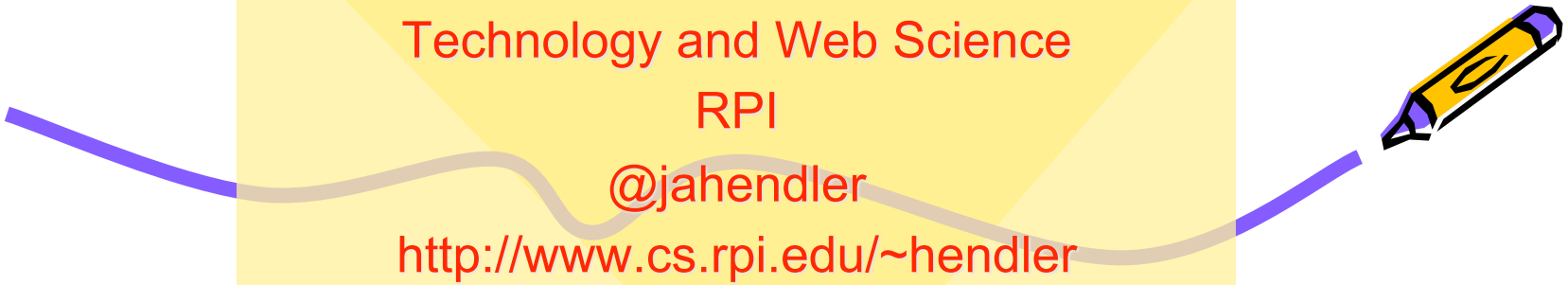
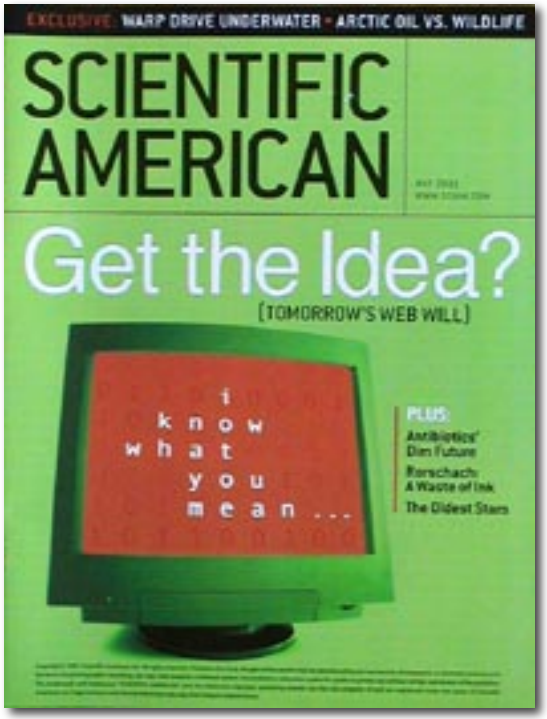
The Semantic Web: Lighter, Faster, Easier

Jim Hendler
Tetherless World Prof. of Information Technology and Web Science
RPI
@jahendler
http://www.cs.rpi.edu/~hendler



The Semantic Web (ca. 2001)



Scientific American Article notes

[Joint starting place:]

I. Semantic Web Vision (TBL)

II. What are the enablers? (in sequence)
Screen Scraping (Ora and TBL)
Data on Web (Ora and TBL)
Zip code link between Data Bases (TBL)
Ontology Independence (JAH)

Effect of Scale (TBL)

"Then, a miracle occurs"

III. What can you do with it? (not necessarily in sequence)
Self-describing documents (JAH)
Logic to encode… (TBL)
Services and Advertising (Ora)
Devices (Ora)
Digital Signatures, Authentication, and Trust (TBL)

(Berners-Lee, Hendler, Lassila; 2001)

Rensselaer

Tetherless World

Semantic Web ca. 2009

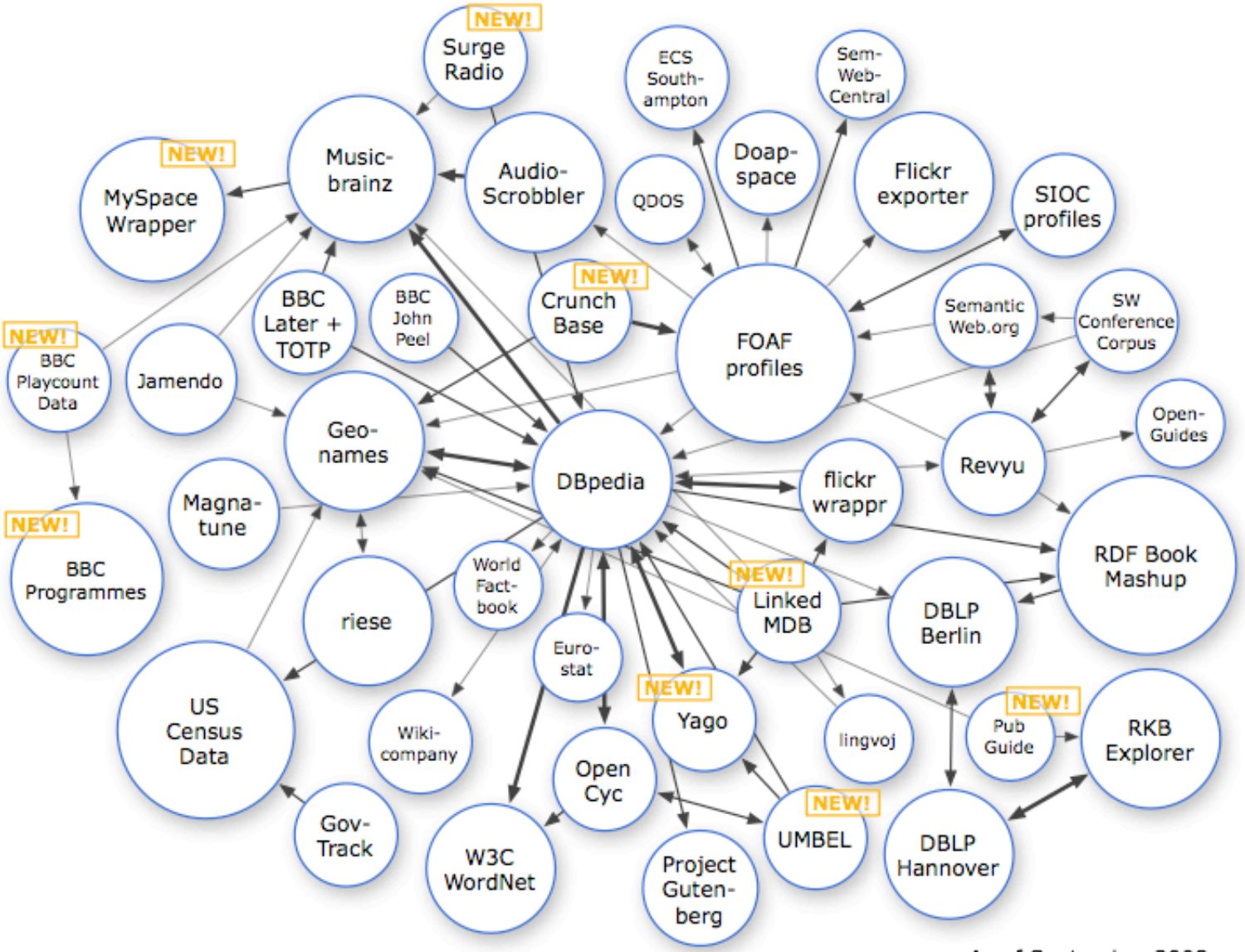
- Semantic Web finding success even in tough market
  - Lots of small companies in the market: Altova… Zepheira (eg. C&P, Franz, Intellidimension, Intellisophic, Ontology Works, Siderean, SandPiper, SiberLogic, TopQuadrant …)
  - Web 3.0 new buzzword: Garlik, Twine, Freebase, Bintro, Siri, Talis, …
  - Semantic Search taking off - Powerset bought by Microsoft for over $100,000,000, hakia, bing, …
- Bigger players buying in
  - 2009 announcements at SemTech (June): Google, New York Times, Oracle, IBM, Yahoo, MS Live Labs, Siri, …
  - 2008: Gartner identifies Corporate Semantic Web as one of three "High impact" Web technologies
  - Tool market forming: AllegroGraph, TopBraid, Pellet2, …
- Government projects in and across agencies
  - Recent open data announcements by UK and US
  - Projects/demos in EU, Japan, Korea, China, India…
  - SKOS update in govt (and private) libraries
- 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, …

Different terms being used in different ways

Rensselaer

Tetherless World

Linked Data: Mashup the data…



Tetherless World

Then build your app on top

Dbpedia mobile

HealthFinland

Semaplorer

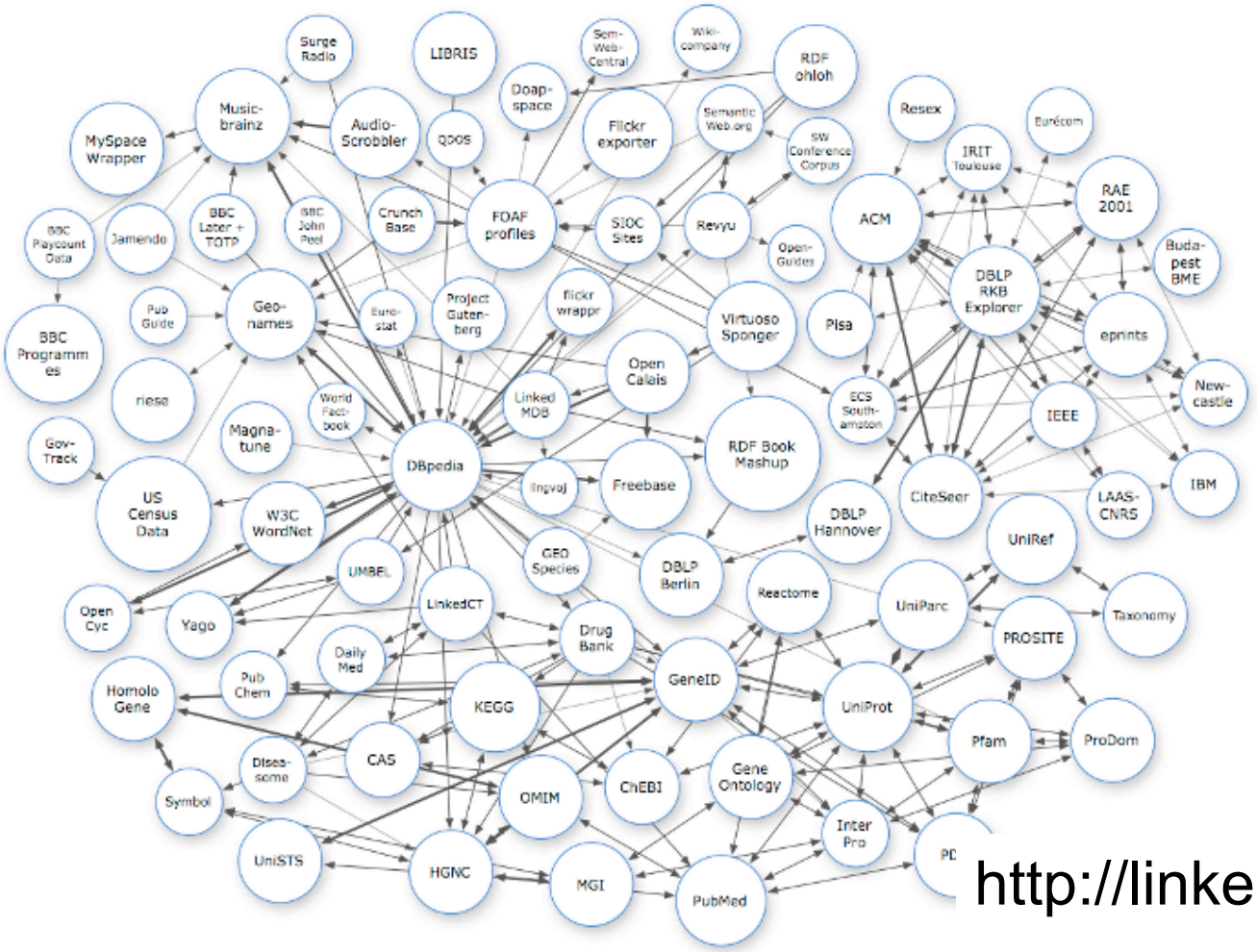
Rensselaer

Tetherless World

Builds over RDF DBs

- The Profile Manager enables you to store information about users and services. It is a Resource Description Framework (RDF) data store and is general nature, so you can store any information that is required by your system. … There are two main benefits offered by a profile store that has been created by using RDF. The first is that RDF enables you to store data in a flexible schema so you can store additional types of information that you might have been unaware of when you originally designed the schema. The second is that **it helps you to create Web-like relationships between data, which is not easily done in a typical relational database.**

http://msdn2.microsoft.com/en-us/library/aa303446.aspx - 12/06



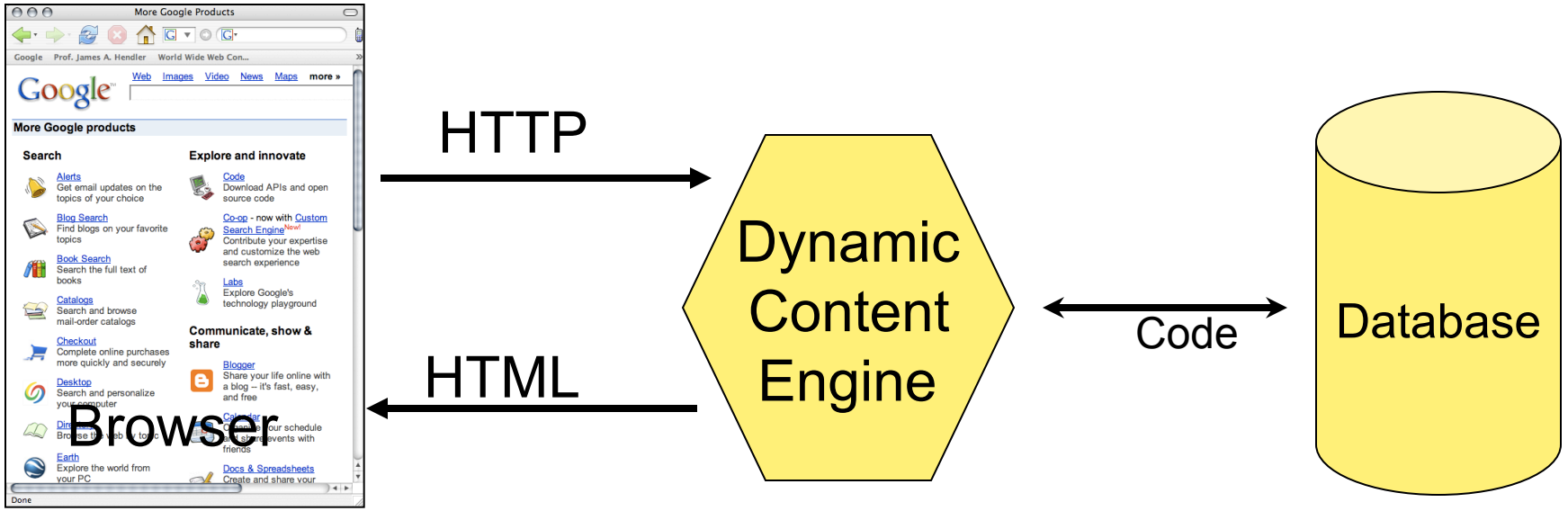
http://linkeddata.org/

The linked open data cloud now has tens of billions of assertions, and new sources are being added rapidly

Rensselaer

Tetherless World

Traditional Web applications



Rensselaer

Tetherless World

Semantic Web applications

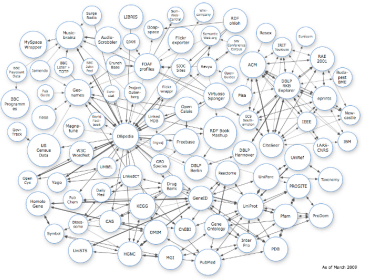
Browser

HTTP (Sparql)

HTML

Dynamic Content Engine

Code



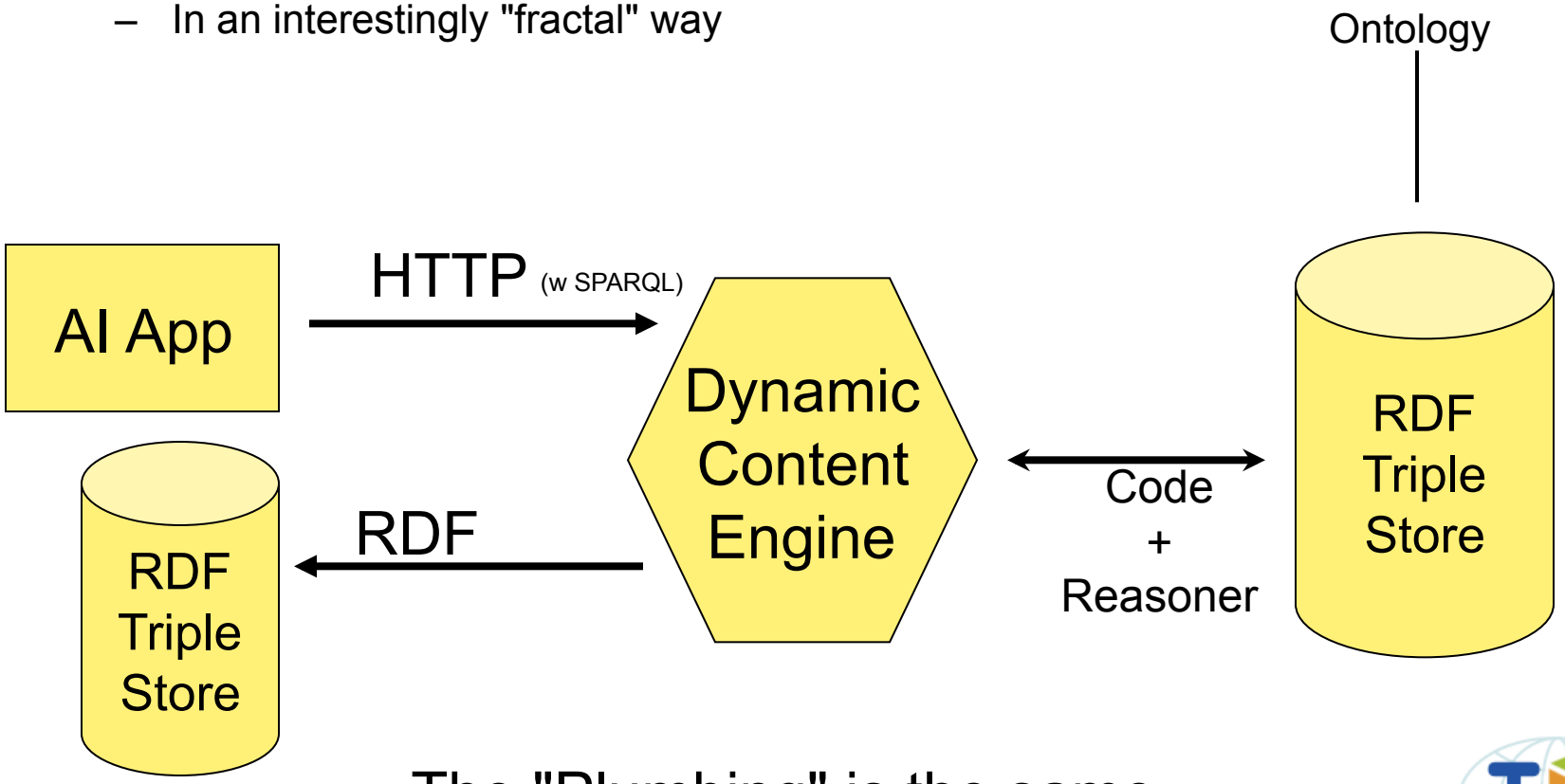
Do your mashup on the underlying data instead of presentations thereof

Rensselaer

Tetherless World

Semantic Web applications

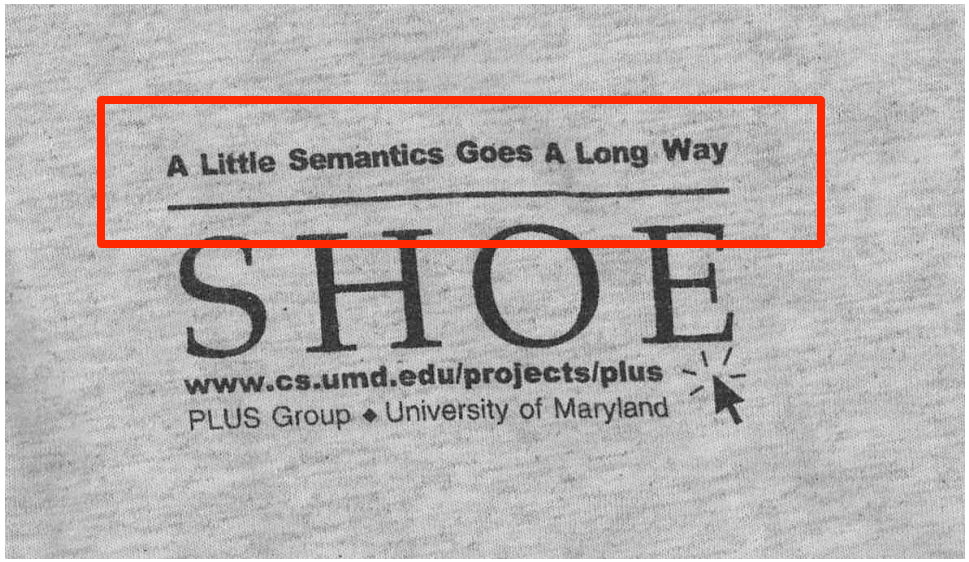
- And a similar model can power the "high end" Semantic Web applications
  - In an interestingly "fractal" way



The "Plumbing" is the same

Linked Data + Semantics

- "Linked Data" 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!**

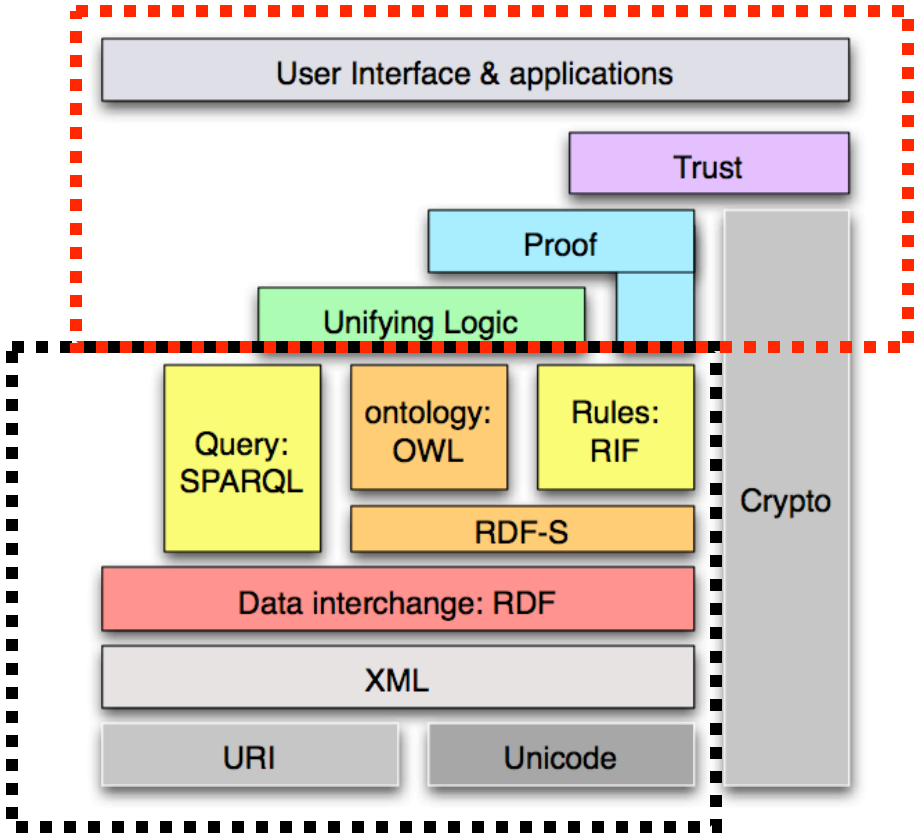


http://www.cs.rpi.edu/~hendler/LittleSemanticsWeb.html

Rensselaer

Tetherless World

Semantic Web "Stack"



Research Activities

Standards activities

ca. 2006 (Tim Berners-Lee)

Rensselaer

Tetherless World

A myth that needs debunking

- ***The Semantic Web needs Ontologies (true)***
- ***But Ontologies*** are
  - Inefficient (slow)
  - Complicated to express (Heavy)
  - Difficult to Build (Hard)

  (false)
- We can build them:
  - ***Faster, Lighter, Easier!!***

Rensselaer

Tetherless World

Traditional AI ontology

- *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

Sem Web use case

- *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 in many unexpected contexts
- FOAF
  - 10s of millions of Foaf people
    - (not necessarily distinct individuals)
  - Exported by a growing number of providers
    - If you use LiveJournal, you have a FOAF file
      - Also flickr, ecademy, tribe, joost, …
      - Apps to export Foaf from Facebook and other soc netw sites
  - Becoming de facto standard for open social networking

A lot more users than the NCI ontology!

Why?

- NCBI view: 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)
  - But still not "Web Scale"

In this view OWL is a formal *knowledge representation* standard

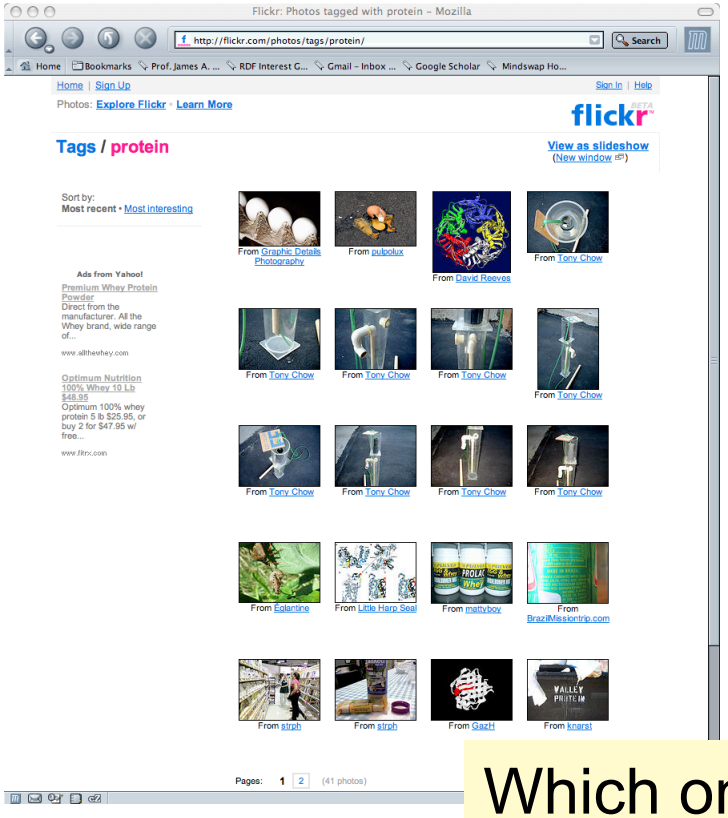
O*ntology*: the traditional view

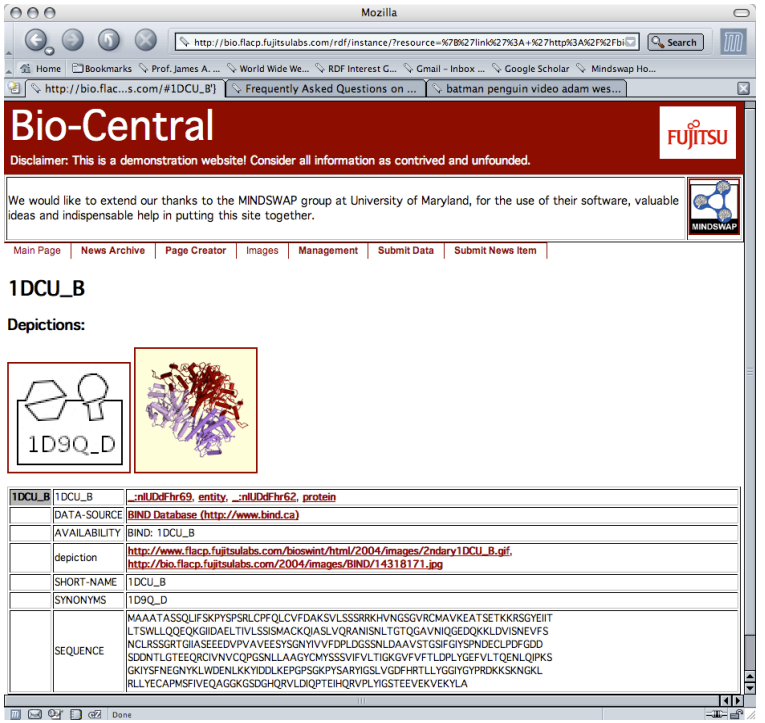
- **O**ntology as Barad-Dur (Sauron's tower):
  - Extremely powerful!
  - ~~Patrolled by Orcs~~ Decidable Logic basis
    - Let one little ~~hobbit~~ inconsistency in, and the whole thing could come crashing down

Rensselaer

Tetherless World

The argument for this seems compelling





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

Rensselaer

Tetherless World

But the cost is high

- 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**
- **The modeling is very expensive and the return on investment must be very high!**

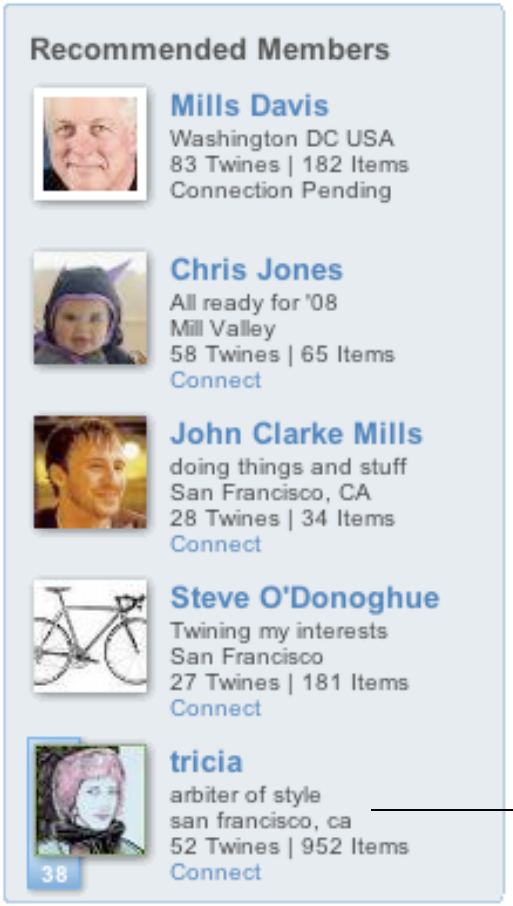
Analogy: the pre-Web hypertext book

A better alternative for Web Development

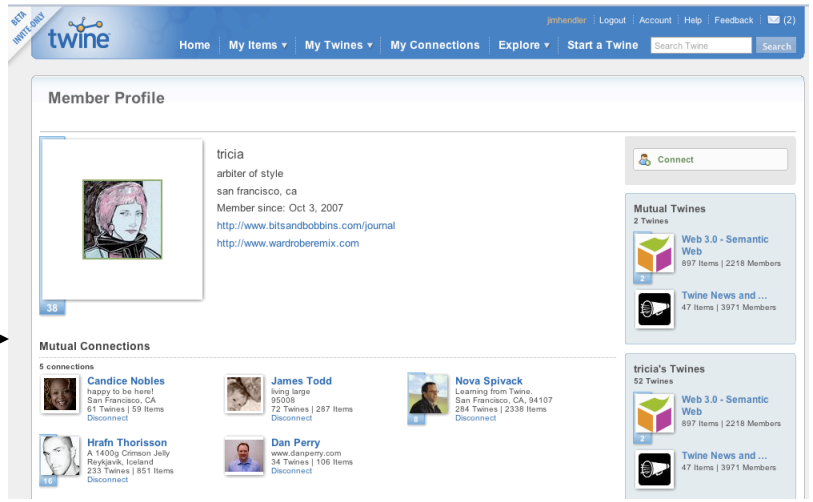
- RDFS and OWL are based on RDF, a language designed for the (Semantic) Web
  - Built with Web architecture in mind
    - Exploits Web infrastructure, respects W3C TAG recommendations
      - Internationalization, accessibility, extensibility
  - Fits the ***Web culture***
    - Open and extensible, 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"

Analogy: HTML

Very simple "reasoning"



- 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 so "first" is more important than "there somewhere"



Rensselaer

Tetherless World

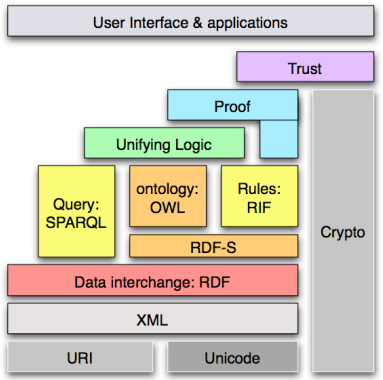
Ontologies?

- Mostly reuse of a few simple ones (Dbpedia terms, foaf, doap, etc.)
  - Faster
- Uses simple parts of language (RDFS and a very small amount of OWL)
  - Lighter (sometimes called "lightweight ontologies"
- Mostly small and "local"
  - Easier

Reasoning?

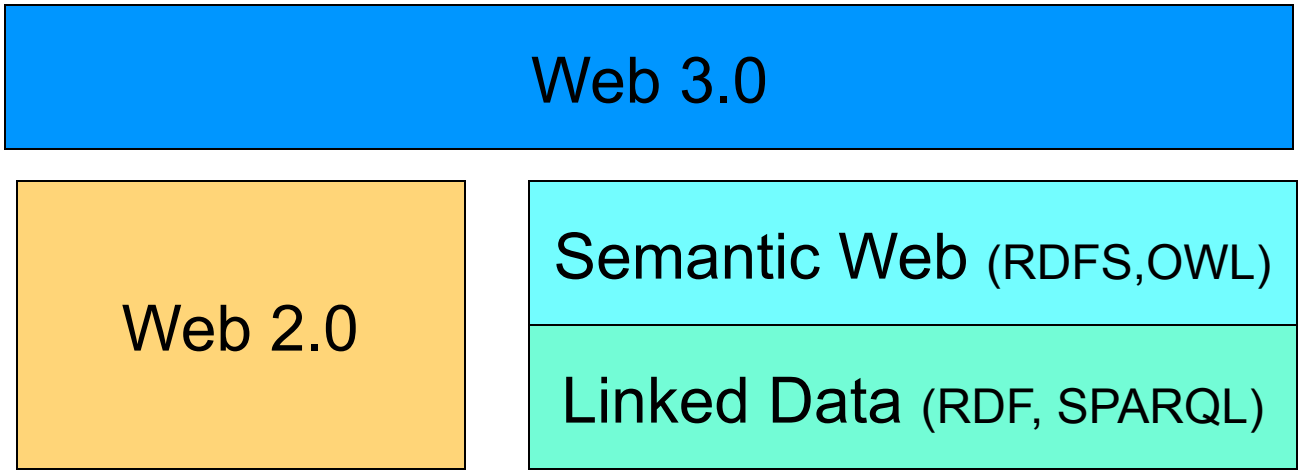
- Very little
  - Mainly just which data in one sphere is related to another
    - (easy)
  - Mainly based on small vocabularies
    - (Light)
  - Mainly procedural
    - (fast)

Evolving standards



- **SPARQL: Query language for (distributed) triple stores**
  - **the "SQL of the Semantic Web"**
- GRDDL/**RDFa**: Integration of HTML and Semantic Web
  - "embedding" RDF-based annotation on traditional Web pages
  - Both Yahoo! and Google now supporting RDFa
- OWL 2.0: New features, specialized subsets
  - **OWL RL – simplification, identity, scaling to large datasets**
- RIF: Rules Interchange Format
  - representing rules on the Web
  - linking rule-based systems together
- And more…
  - **SKOS** thesaurus standard
  - Multimedia annotation, Web-page metadata annotation, Health Care and Life Sciences (LSID), privacy, Sem Web Service, etc.

The new meme: Web 3.0



Web 3.0 extends current Web applications using Semantic Web technologies and graph-based, open data.

Rensselaer

Tetherless World

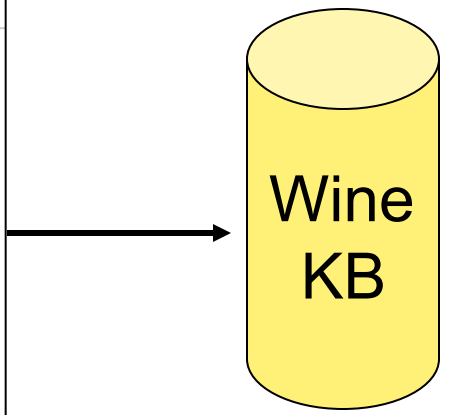
The making of a 3.0 app…

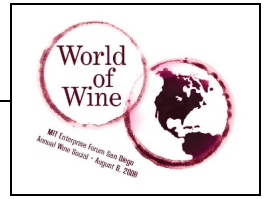
- The Wine Agent ca '85 : Reasoner with knowledge of wine and food pairings

Swordfish subclass BlandFish subclass Fish
⇒ Dry + medium-body + White
⇒ Drink EdenValleyChardonnay

  - Used the wines in a particular wine cellar
    - Hard wired in
  - Eventually completed with "correct" wine recommendations
    - You disagree, tough! You're wrong.

Wine Agent 2007





Rensselaer

Recommendation Wiki (Michaelis, 07)

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 ⊗			

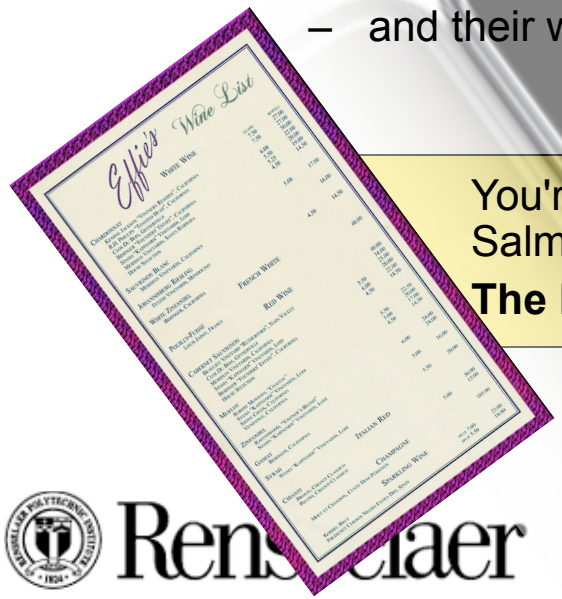
back

Rensselaer

Tetherless World

Wine Agent 3.0

- Coming soon: Add the Data!
  - Phone knows your location (and thus what restaurant you are in)
    - And the menu
    - And the wine list
  - Phone knows who else is there
    - Your Facebook network is there with you
    - So are other people with the application
  - and their wine preferences



You're having sole, Jane beef, and Fred the Salmon Special

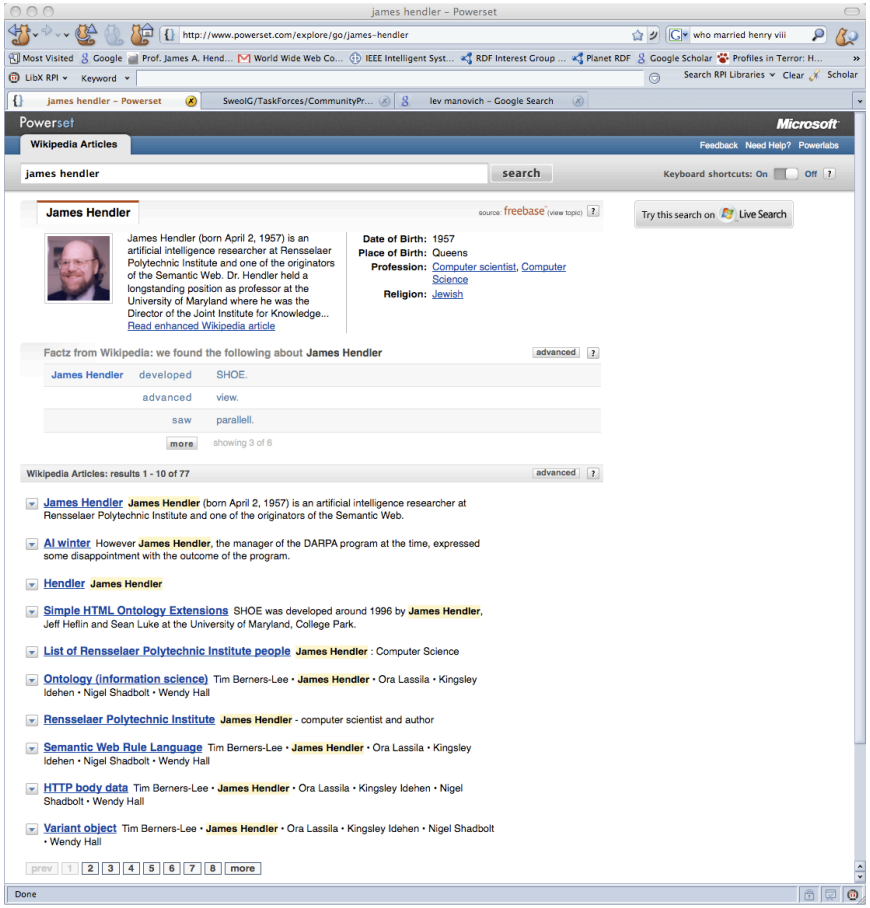
The Flowers Pinot would be a great choice!

Patton 08

Rensselaer

Tetherless World

Web 3.0 examples

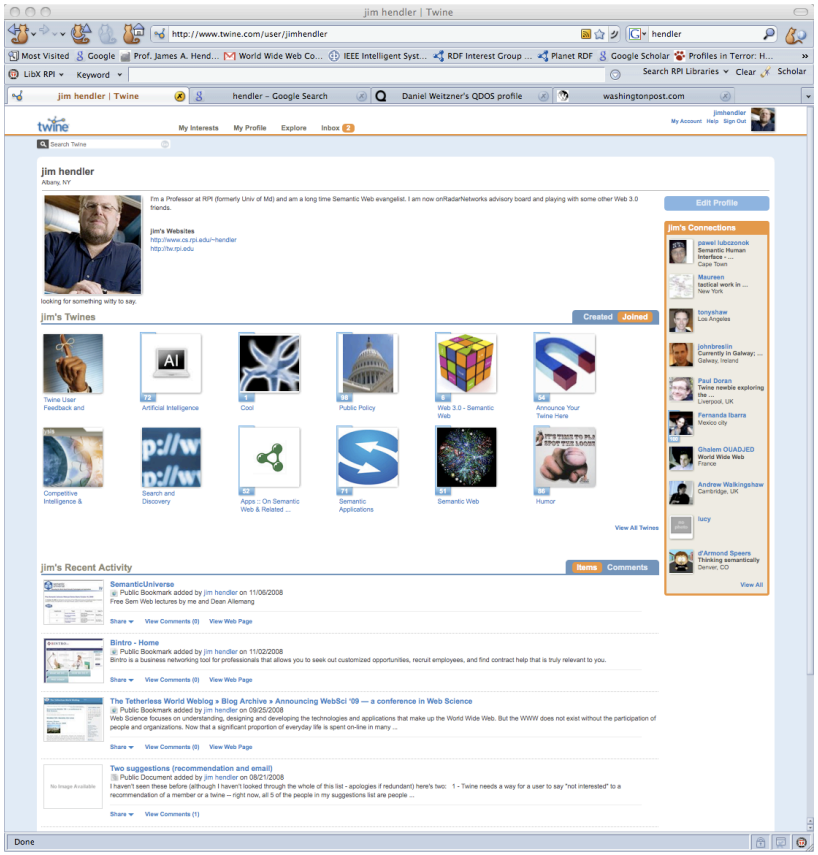


Semantic Search (Powerset.com)

Rensselaer

Tetherless World

Web 3.0 examples

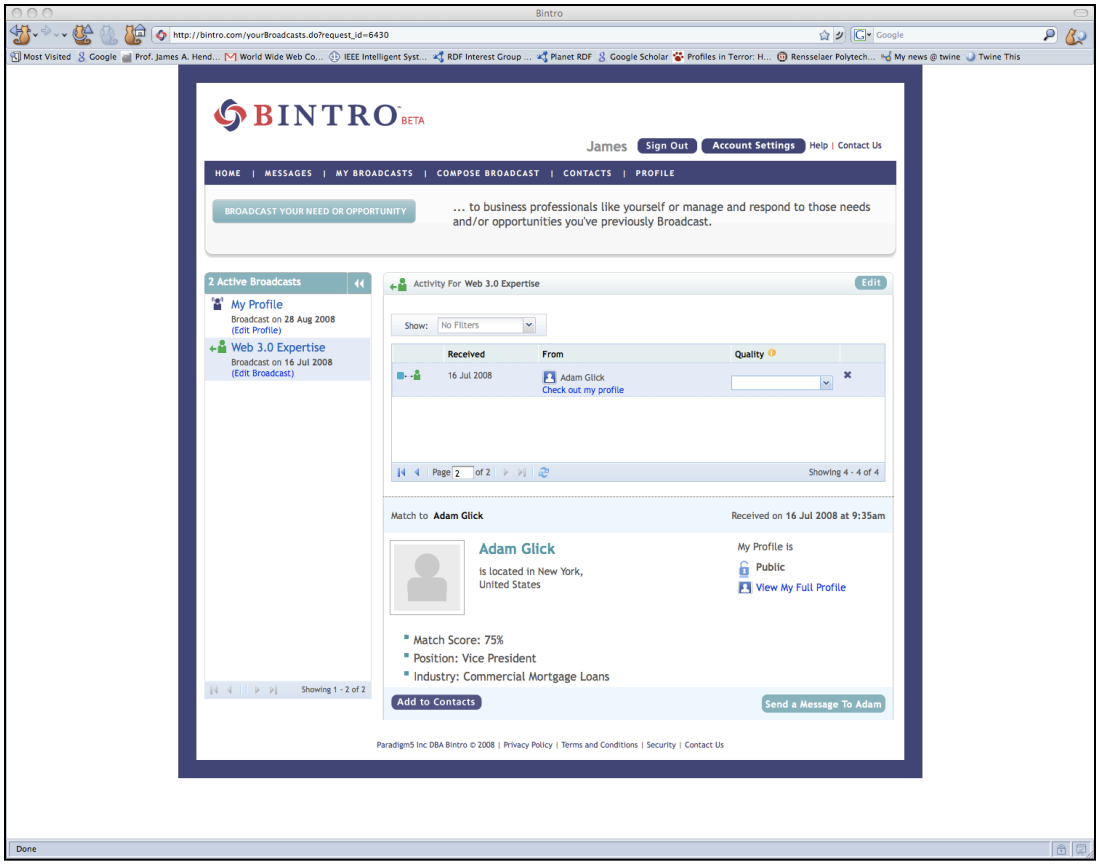


Enhanced Social Networks (twine.com)

Rensselaer

Tetherless World

Web 3.0 examples

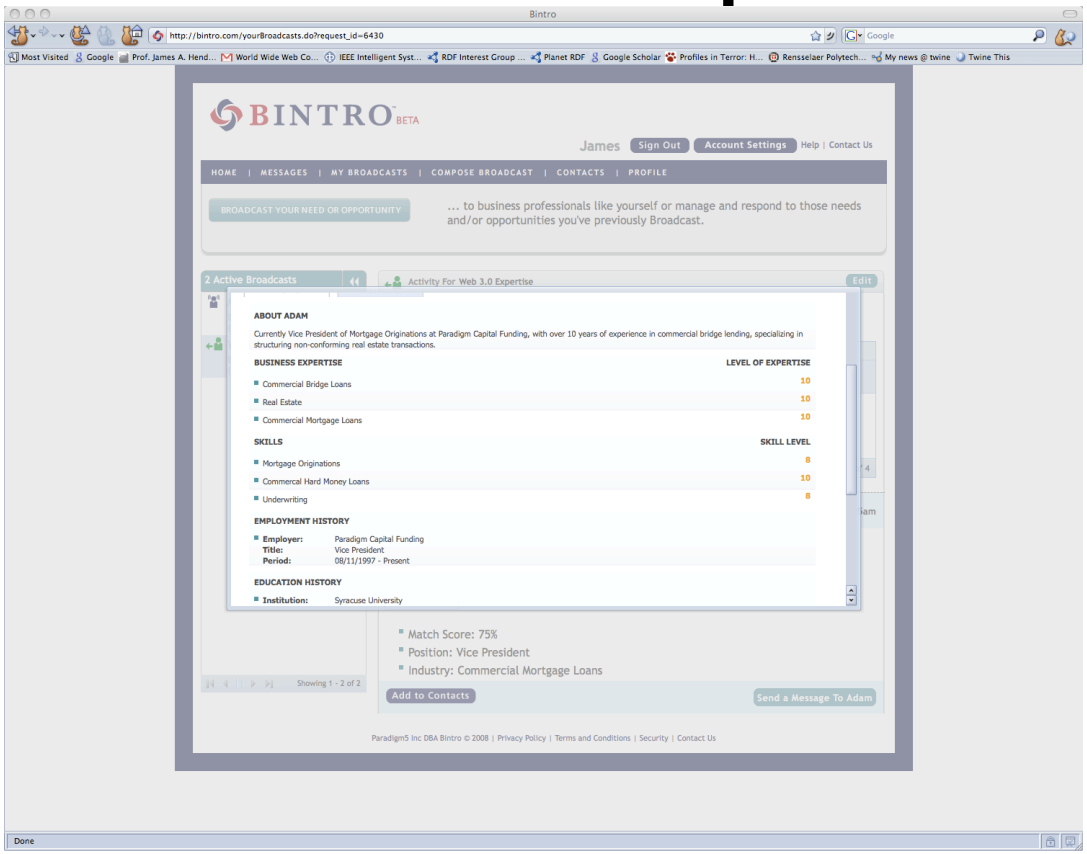


Semantic Match (bintro.com)

Rensselaer

Tetherless World

Web 3.0 examples

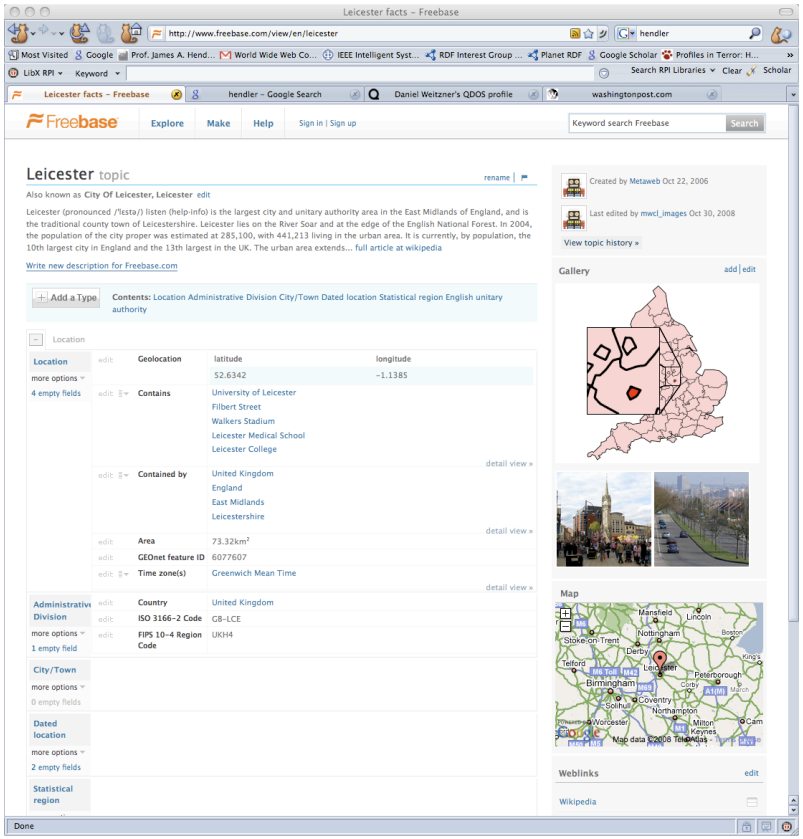


Semantic Match (bintro.com)

Rensselaer

Tetherless World

Web 3.0 examples



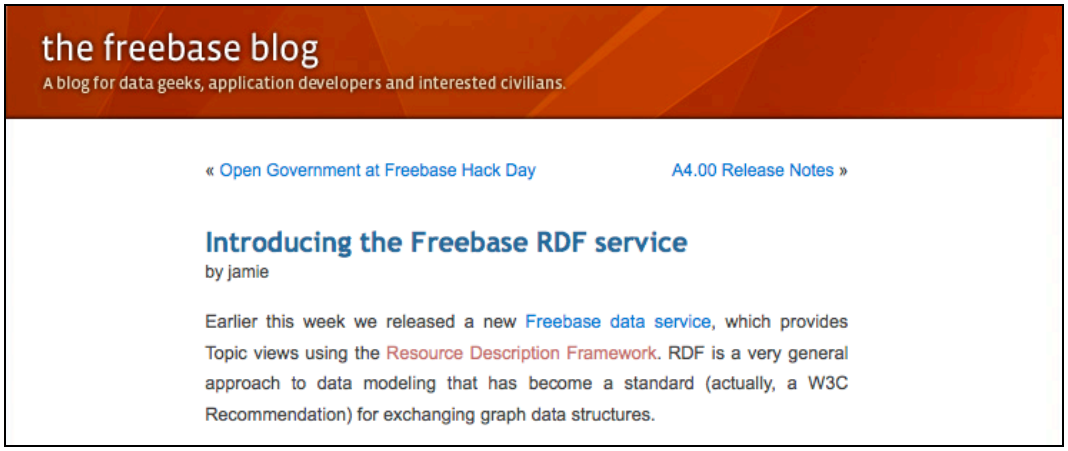
Social database (freebase.com)

Rensselaer

Tetherless World

Web 3.0 - RDF APIs

- RDF also starting to provide interoperability between Web applications in Web 2.0 and Web 3.0
  - Many Web 2.0 apps already can dump RDF
    - Flickr, mySpace, facebook, livejournal…
  - Web 3.0 apps are doing so as well

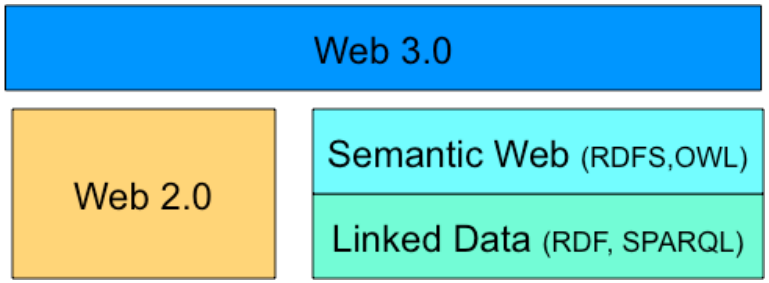


Rensselaer

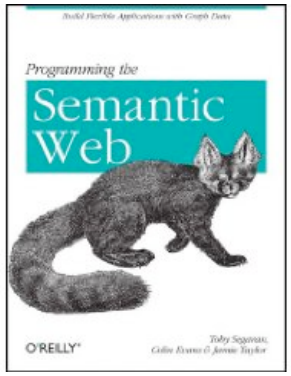
Tetherless World

Web 3.0 excitement

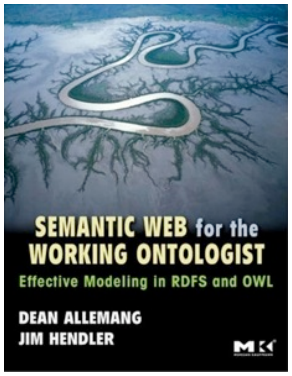
- Significant and growing commercial interest…
  - Web: Google
  - Web 2.0: Facebook, Wikipedia, YouTube, Flickr, …
  - Web 3.0: the big one is still out there



How can I learn more?



http://www.amazon.com/Programming-Semantic-Web-Toby-Segaran/dp/0596153813



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

Rensselaer

Tetherless World

Bottom line

- The Semantic Web, powered by technologies such as RDFS, SPARQL, and a little bit of OWL is showing tremendous promise
  - Linked Data – focus on open world and network effect
    - Mashup the data (Web like) and build you app (Web scale)
    - Traditional Web architectures work just fine
  - Web 3.0 – embed the power of the Semantic Web in large scale Web apps
    - Closer to Web 2.0 in look and feel
    - Similar implementation approach

Lighter, Faster, Easier!

WebSci10

- http://websci09.org was a great conference
  - 350 people
  - #websci09 a top-10 twitter trend
- WebSci10 will be help co-located with WWW2010 in Raleigh, NC, US
  - Subscribe to wsri-announce@webscience.org for continued information
- And please submit your work!

Rensselaer

Tetherless World