The Maturing Semantic Web: Lessons in Web-Scale Knowledge Representation

Dr. Mark Greaves
Vulcan Inc.
markg@vulcan.com

What is the Semantic Web? 4 Answers

- **The Web of Data**
  - A fully distributed web-based system to publish logical assertions
  - A way to link to someone else's data, augment it, and add to it
  - Democratic, crowd-based, scalable knowledge engineering
  - The hottest area of Artificial Intelligence right now

What is the Semantic Web? 4 Answers

- **The Web of Data**
  - A fully distributed web-based system to publish logical assertions
  - A way to link to someone else's data, augment it, and add to it
  - Democratic, crowd-based, scalable knowledge engineering
  - The hottest area of Artificial Intelligence right now
- **The set of W3C standards for simple logic languages on the web**
  - RDF, RDFS, OWL (3 versions), OWL2
  - Weaker than First Order Logic, more easily authorable, decidable, tractable in most cases using tableaux provers

What is the Semantic Web? 4 Answers

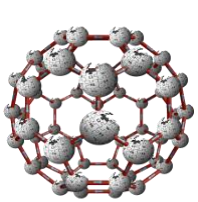
- **The Web of Data**
  - A fully distributed web-based system to publish logical assertions
  - A way to link to someone else's data, augment it, and add to it
  - Democratic, crowd-based, scalable knowledge engineering
  - The hottest area of Artificial Intelligence right now
- **The set of W3C standards for simple logic languages on the web**
  - RDF, RDFS, OWL (3 versions), OWL2
  - Weaker than First Order Logic, more easily authorable, decidable, tractable in most cases using tableaux provers
- **The largest formal knowledge base on Earth**
  - Also the messiest formal knowledge base on Earth

What is the Semantic Web? 4 Answers

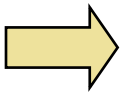
- **The Web of Data**
  - A fully distributed web-based system to publish logical assertions
  - A way to link to someone else's data, augment it, and add to it
  - Democratic, crowd-based, scalable knowledge engineering
  - The hottest area of Artificial Intelligence right now
- **The set of W3C standards for simple logic languages on the web**
  - RDF, RDFS, OWL (3 versions), OWL2
  - Weaker than First Order Logic, more easily authorable, decidable, tractable in most cases using tableaux provers
- **The largest formal knowledge base on Earth**
  - Also the messiest formal knowledge base on Earth
- **A revolution in the way we think of data, crowds, and schemas**
  - Massive, partial, participatory, logically weak, dynamic, schema-last
  - A way to democratize and scale knowledge bases and knowledge systems
  - A route to impact for AI technologies

Talk Outline: The Maturing Semantic Web

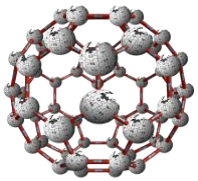
- **The Origins of the Semantic Web**
  - DARPA's DAML Program
  - RDF, OWL, and the Semweb Infrastructure
- **Semantic Web Evolution to 2009**
  - Three Generations of Semantic Dreams
  - Markets and Companies
- **The Fourth Generation**
  - A Scalable Revolution



Talk Outline: The Maturing Semantic Web

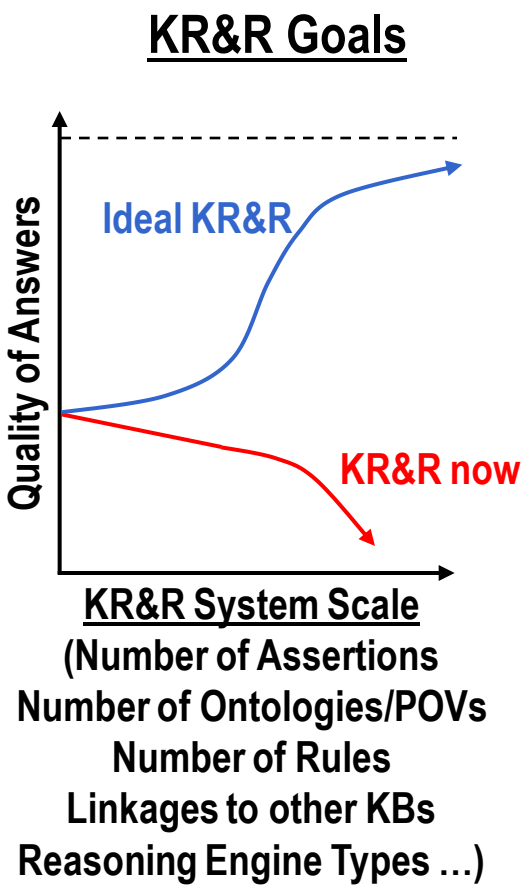


- **The Origins of the Semantic Web**
  - DARPA's DAML Program
  - RDF, OWL, and the Semweb Infrastructure
- **Semantic Web Evolution to 2009**
  - Three Generations of Semantic Dreams
  - Markets and Companies
- **The Fourth Generation**
  - A Scalable Revolution



At the End of the 90s: Traditional KR and the Google Property

- **We seek KR systems that have the "Google Property:"** ***they get (much) better as they get bigger***
  - Google PageRank™ yields better relevance judgments as it indexes more pages
  - 1990's KR&R systems have the antithesis of this property
- **So what are the components of a scalable KR&R system?**
  - Distributed, robust, reliable infrastructure
  - Multiple linked ontologies and points of view
    - Single ontologies are feasible only at the program/agency level
    - Multiple authors and overlapping data sources
    - Private and public knowledge
  - Mixture of deep and shallow knowledge
  - Tractable reasoning algorithms
  - Tolerant KB – you are typically doing open-world reasoning (no NAF), things go away, contradiction is present, data is incomplete and dirty, computing must be resource-aware, surveying the KB is not possible
  - (Relatively) easy for non-KE's to author, validate, and maintain



Scalable KR&R Systems should look just like the Web!!

VULCAN

The Roots of the Semantic Web

- **Semantic technology has been a distinct research field for decades**
  - Symbolic Logic (from Russell and Frege)
  - Knowledge Representation Systems in AI
    - Semantic Networks (Bill Woods, 1975)
    - DARPA and European Commission programs in information integration
    - Development of simple tractable “description logics” for classification
    - Conceptual Graphs and this community
  - Relational Algebras and Schemas in Database Systems
- **Library Science (classifications, thesauri, taxonomies)**
- **What’s new was the Web!**
  - The material needed to answer almost any question is somewhere on the web
  - A massive infrastructure of data servers, protocols, authentication systems, presentation languages, and thin clients that can be leveraged
  - A way around needing the “big data warehouse”

The Beginnings of the US Semantic Web: DARPA's DAML Program

Problem:

Computers cannot process most of the information stored on web pages

Solution:

Augment the web to link machine-readable knowledge to web pages

- Extend RDF with Description Logic
- Extensibility via frame-based language design
- Create the first fully distributed web-scale knowledge base out of networks of hyperlinked facts and data

Approach:

Design a family of new web languages

- Basic knowledge representation (OWL)
- Reasoning (SWRL, OWL/P, OWL/T)
- Process representation (OWL/S)

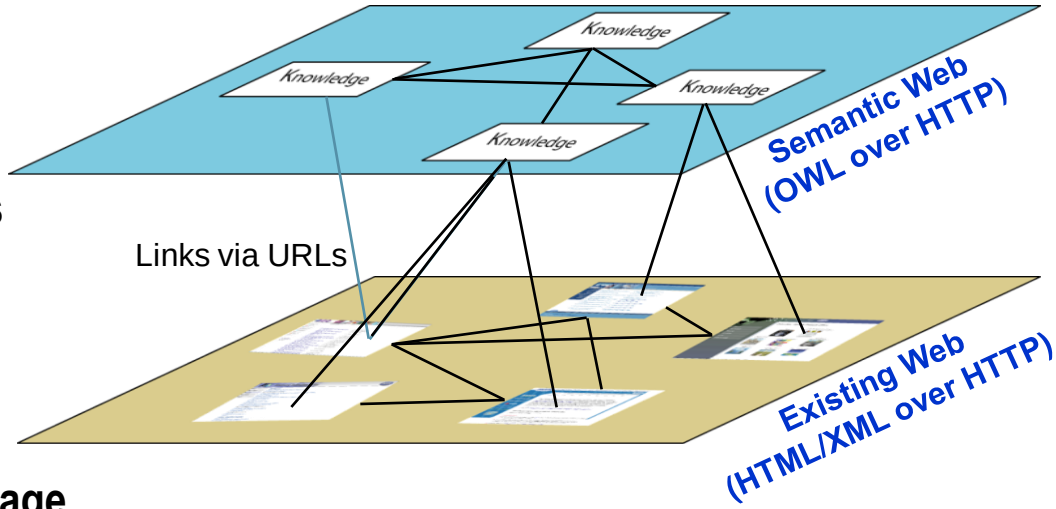
Build definition and markup tools

Link new knowledge to existing web page elements

Test design approach with operational pilots in US Government

Partner with parallel EU efforts to standardize the new web languages

Computers require explicit knowledge to reason with web pages

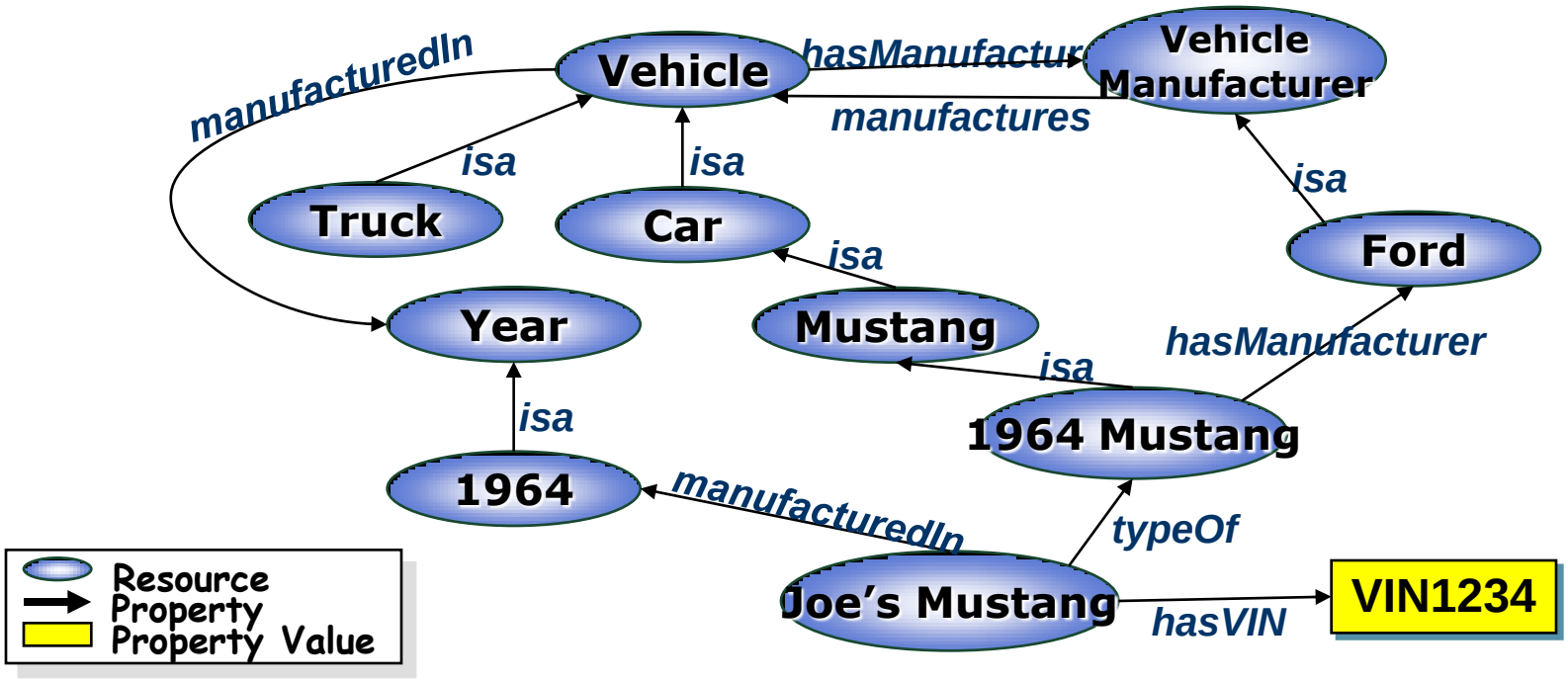


People use implicit knowledge to reason with web pages

VULCAN

What is RDF?

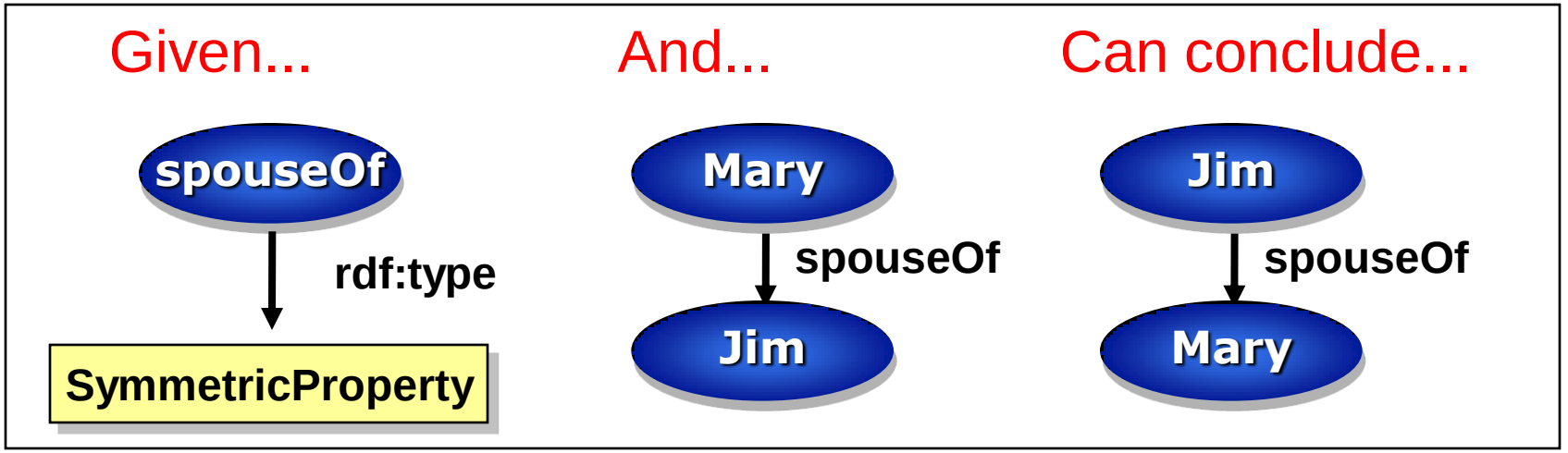
- **Defines the terms used to describe and represent an area of knowledge, using web-friendly technologies**
  - Specified by triples (resource, property, resource) or (resource, property, value)
  - Precise enough to be interpreted by machines
  - Enables reuse of domain knowledge; makes domain assumptions explicit



VULCAN

What Does OWL Add?

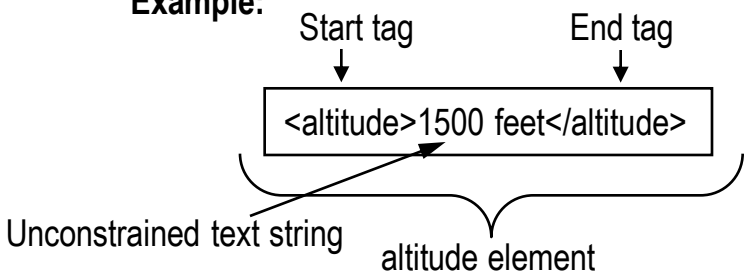
- **More Semantic Expressiveness**
  - Relations between classes
    - Equiv. Class (e.g., US_President and PrincipalResidentofWhiteHouse)
    - Disjoint Class (e.g., Male and Female)
  - Complex Classes (intersectionOf, unionOf, complementOf)
  - Property characteristics (inverseOf, transitive, symmetric, etc.)
  - Cardinality constraints (e.g., birthMother has exactly one value)
- **Ability to combine facts and make inferences**



From XML to OWL
Increasingly Expressive Options for Web Data Markup

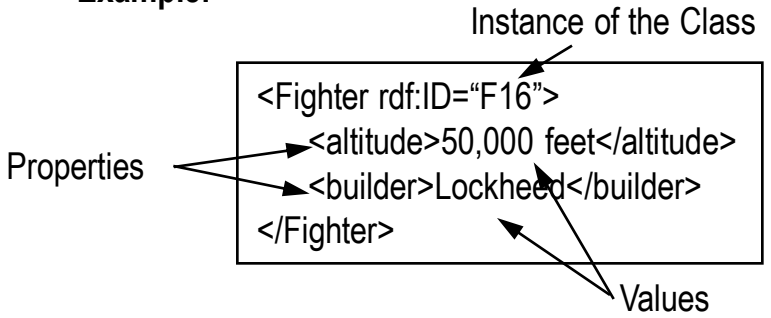
XML

Issue addressed: how to express data in text?
XML Solution: "wrap" data within start tag/end tags, and empower users to create their own tags
Example:



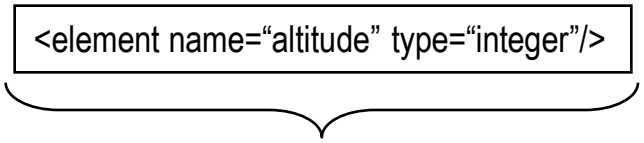
RDF and RDF Schema

Issue addressed: how can data support statements?
RDF Solution: use a *subject, property, object* pattern
Example:



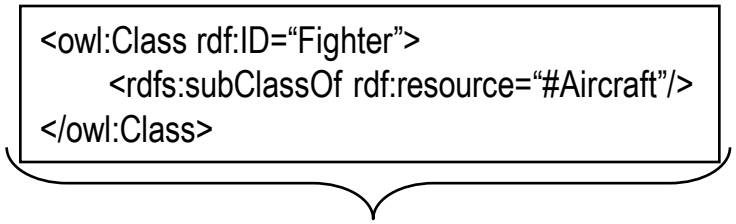
XML Schema (XMLS)

Issue addressed: how should the type structure of the data be expressed?
XML Schema Solution: XML templates
Example:



OWL

Issue addressed: how to express data semantics?
OWL Solution: use inheritance and a description logic to express restrictions and describe entailment
Example:



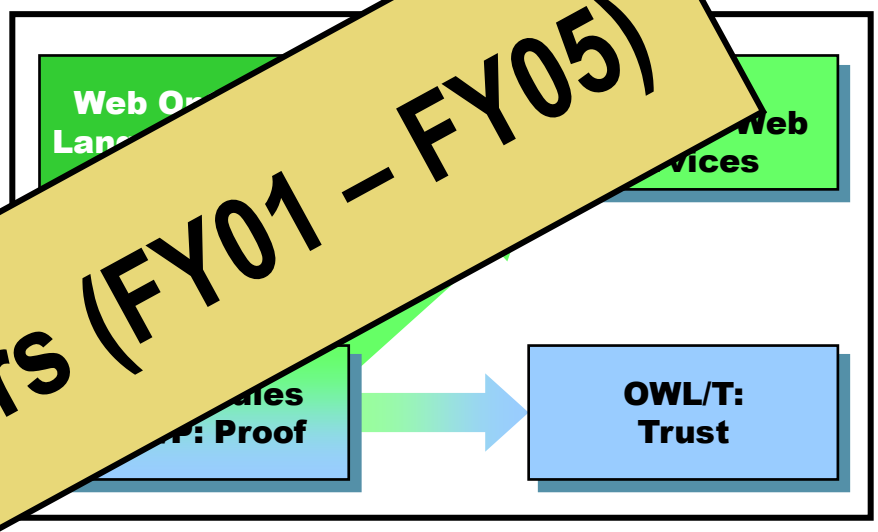
HTML → XML & XMLS → RDF → OWL

VULCAN

DAML Program Elements

- **Web Ontology Language (OWL) (2/10/04)**
  - ✓ – Enables knowledge representation and tractable inference in a web standard format
  - ✓ – Based on Description Logics and RDF
- **OWL Reasoning Languages**
  - ✓ – SWRL and SWRL-FOL: Supports business rules, policies, and linking between distinct OWL ontologies
  - – OWL/P Proof Language: Allows softwar components to exchange chains of
  - – OWL/T Trust Language: Re OWL and SWRL infere
- **Semantic Web**
  - ✓ – Allows ions
  - ✓ – WL) with d shows how to web services
- **OWL To**
  - ✓ – www.semwebcentral.org and www.daml.org
- ✓ **Several US Govt pilots and prototypes**

~$45M over 5 years (FY01 – FY05)



Completed standards process

Started standards process

Unfinished

Each DAML Program Element includes specifications, software tools, coordination teams, and use cases

VULCAN

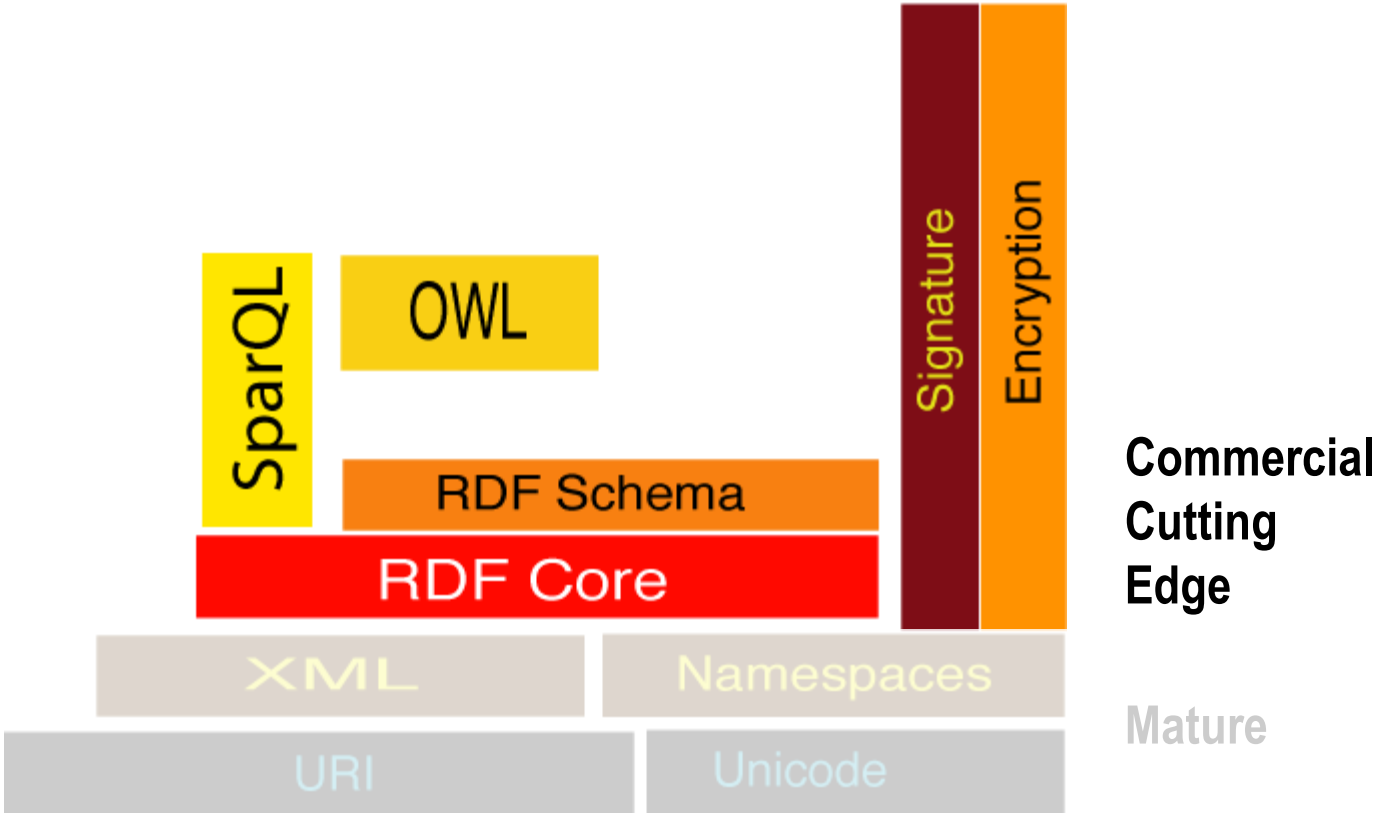
The Semantic Web in 2009

Mature

"The Famous Semantic Web Technology Stack"

VULCAN

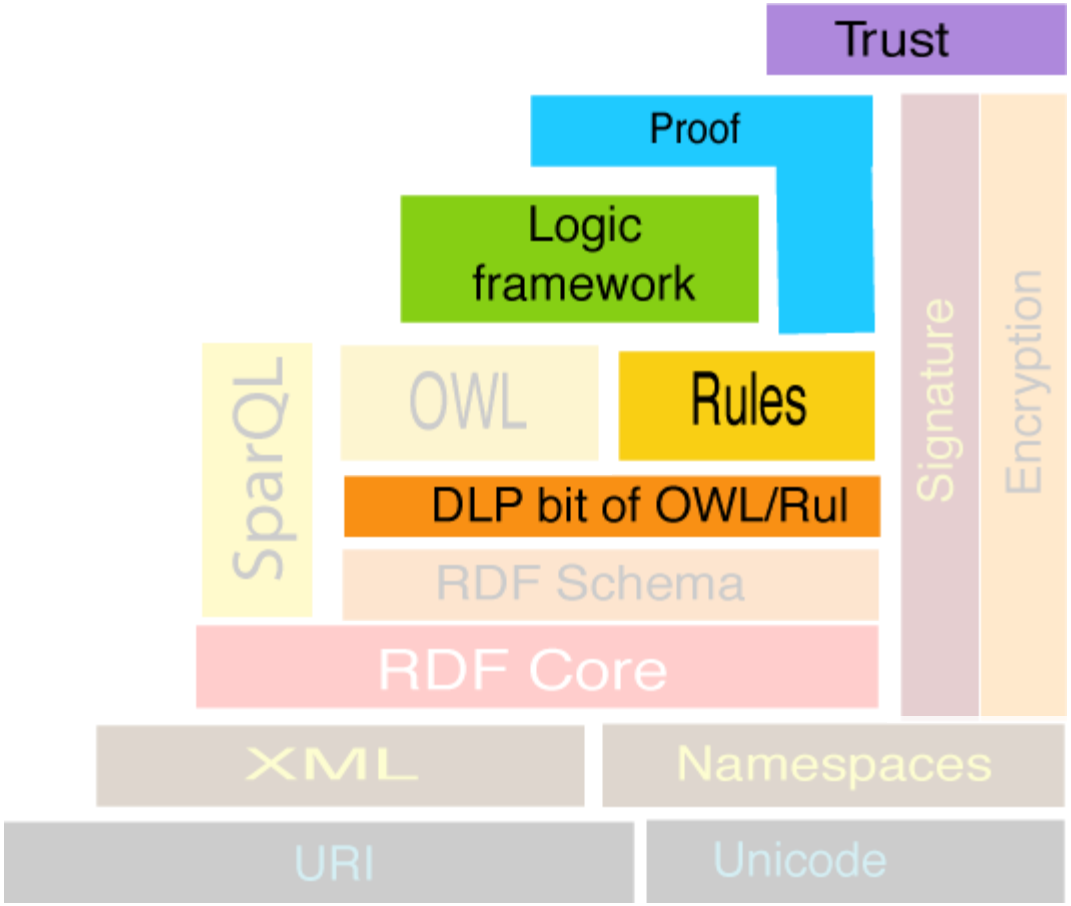
The Semantic Web in 2009



"The Famous Semantic Web Technology Stack"

VULCAN

The Semantic Web in 2009



Active Research and Standards Activity

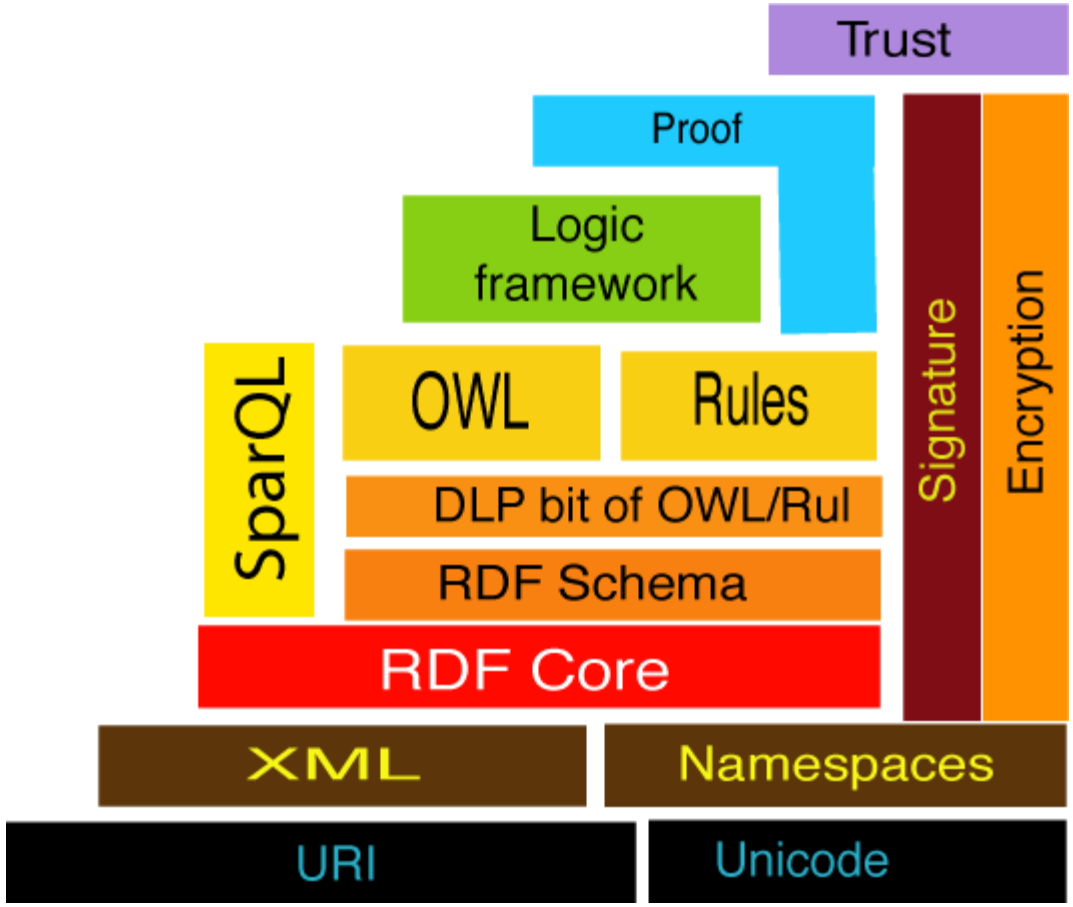
Commercial
Cutting
Edge

Mature

"The Famous Semantic Web Technology Stack"

VULCAN

The Semantic Web in 2009



Active Research and Standards Activity

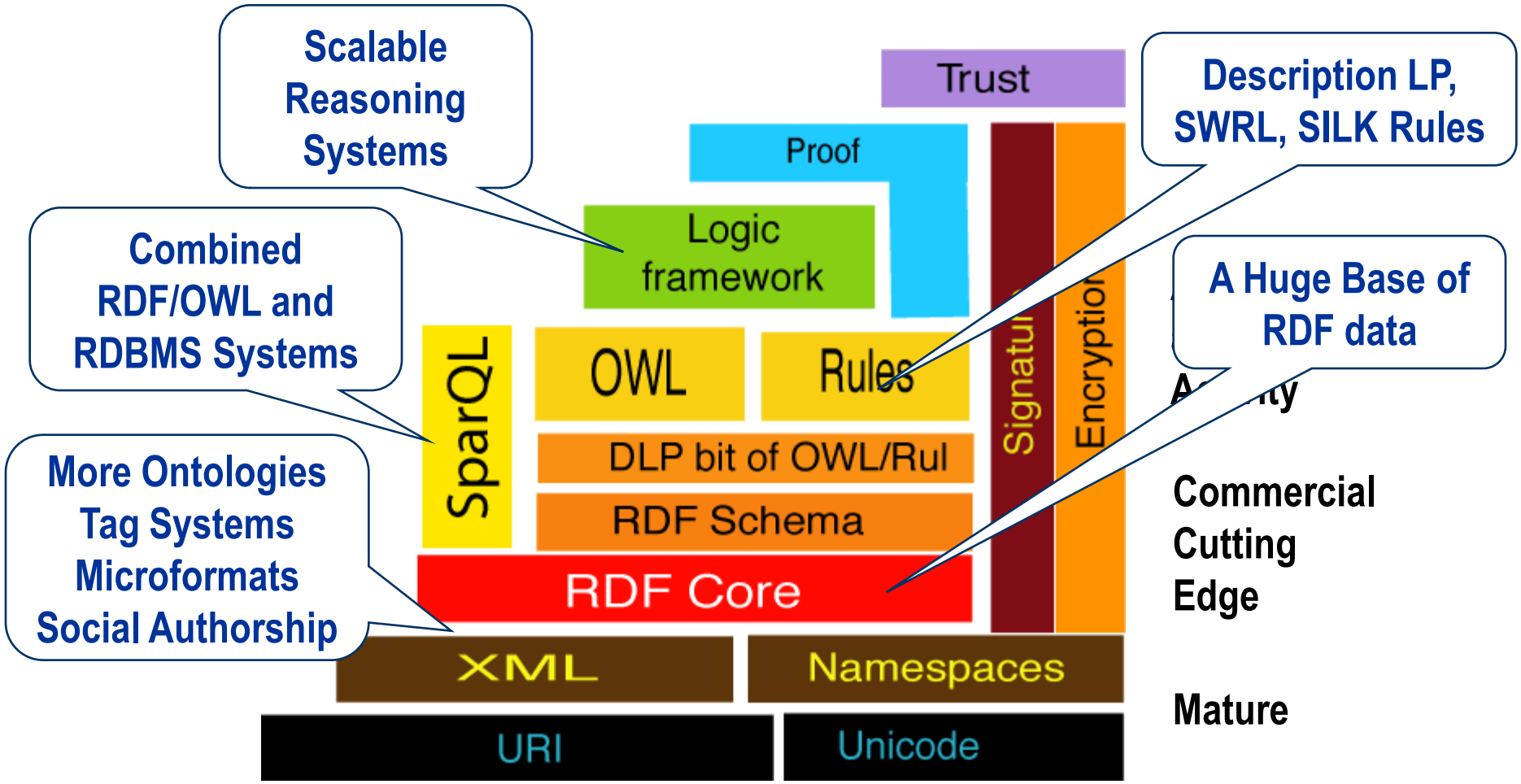
Commercial Cutting Edge

Mature

"The Famous Semantic Web Technology Stack"

VULCAN

Completing the Semantic Web Picture



Other Technologies Impact the Semantic Web

VULCAN

Beyond RDF and OWL: 2009 Semantic Web Infrastructure

Server Infrastructure

- **Markup Languages**
  - HTML-friendly markup dialects: Microformats and RDFa
  - OWL 2 is a Candidate Recommendation
- **Triplestores and SPARQL Servers**
  - Stores for 1B triples now available, though with caveats around write performance
  - Commercial: AllegroGraph, Virtuoso, BigOWLIM, Oracle 11g Semantic Technologies...
  - Open Source: 4Store, Sesame, Redland...
  - Next step is parallel web delivery architectures
- **Entity Name Service (Okkam, DBpedia)**
- **Semantic Web Reasoners**
  - Commercial: Oracle 11g RDFS/OWL engine, Ontobroker, Ontotext, RacerPro
  - Open Source: Pellet, FaCT++...
  - RIF is at W3C Last Call status

User-layer Tools

- **Vocabularies and Design Tools**
  - Ontologies: Dublin Core, FOAF, SIOC...
  - OpenSource: Protégé, SWOOP...
  - Commercial: TopBraid Composer, Knoodl
- **Semweb Data Generation**
  - RDF / RDBMS front-ends
  - NLP parsers into OWL
  - Zemanta-type blogger's assitants
  - Semantic wikis
- **Semweb Data Exploitation**
  - Semweb search engines (Sindice, Watson, Falcon...)
  - Yahoo SearchMonkey / Google Rich Snippets
  - Browser extensions and facets
- **Visualization Tools**
  - Simile Project (http://simile.mit.edu/)
  - Several Commercial Companies

State of Semantic Web Work in the US

- **DAML finished in 2005, with no followons**
  - NIH (Protégé, NCBO), NSF, some small DoD funding
  - PAL/CALO funded broader semantic/AI work

- **But... leading-edge Venture Capital moved in**
  - Vulcan, Crosslink, In-Q-Tel, Benchmark, Intel Capital...

- **An emerging commercialization ecosystem**
  - Startup/Small: Radar, Metaweb, Evri, AdaptiveBlue...
  - Midsized: Metatomix, Dow Jones, Reuters/OpenCalais, Franz...
  - Large: Yahoo!, Google, Oracle, IBM, HP, Microsoft...
  - Semantic web meetup groups in Silicon Valley, Boston, Seattle...

- **Emphasis is mostly *Semantic* dimension of Semantic Web**
  - That was where the money was
  - RDBMS scale and orientation, powerful analytics for Business Intelligence
  - Centralized workflows for ontology definition and management
  - Use cases surrounding corporate data integration and document markup

Semantic Web Work in the EU

- **Continuing Large Public-Sector Investments**
  - Framework 6 (2002-6) – More than €100M in several different programs
  - Framework 7 (2007-13) – ~€1B/year for information and communications technologies
    - Semantics is more present as a general systems technology
    - Future Internet and Digital Libraries thrusts

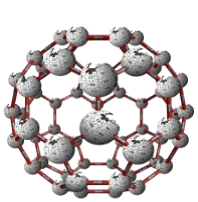
- **Two Dedicated Multi-site R&D Institutes**
  - DERI: 100+ people and the world leader in research
  - Semantic Technology Institute International
  - A strong and growing cadre of graduate students

- **Emphasis on the *Social* and *Web* Dimensions of the Semantic Web**
  - Web-scale Linked Data, social networks, simple scalable imperfect inference
  - Ontology and data dynamics, imperfections, versioning
  - Semantically-boosted collaboration with limited knowledge engineer involvement

Clear R&D leadership but lags in commercialization

Talk Outline: The Maturing Semantic Web

- **The Origins of the Semantic Web**
  - DARPA's DAML Program
  - RDF, OWL, and the Semweb Infrastructure
- **Semantic Web Evolution to 2009**
  - Three Generations of Semantic Dreams
  - Markets and Companies
- **The Fourth Generation**
  - A Scalable Revolution



Evolving Conceptions for the Semantic Web

Initial Semantic Web Conception*

- **Semantic markup would be tightly associated with individual web pages**
  - "Translate the Web for machines"
  - RDFa shows this is still a powerful vision
- **Core problem is labeling free-text web pages with a (pre-defined) ontology markup vocabulary**
  - Entity extraction and other lightweight NLP
  - Document segmentation technologies
  - Manual annotation
- **Need an all-encompassing ontology or set of logically compatible ontologies**
- **Small number of knowledge engineers do semantic annotation because the modeling problems are so hard**
  - Knowledge engineers rarely get markup right because they aren't domain experts

The Semantic Web in 2009

- **The Web is a publishing platform for formal knowledge as well as pages**
  - Semantic data doesn't have to be associated with HTML web text (just a URI)
  - Huge numbers of knowledge publishers
  - Simple RDF and `owl:sameAs` links
- **Core problem is maintaining a set of evolving and partial agreements on semantic models and labels**
  - Consensus is a human social problem
  - There will be massive numbers of overlapping ontologies and class hierarchies, and lots of bad data
  - Hard problem is cost-effectively maintaining semantic models and labeling data
- **Supplemental semantics is carried in the free-text web**

* By most people but not Tim Berners-Lee

First Generation Semantic Web Applications

Semantically-Boosted Search and Classification

- **A really old problem type**
  - Semantics as the keystone technology for unstructured Information Retrieval
  - Requires powerful NLP and document interpretation systems
    - Often also requires powerful semantic representations (e.g., events or causality)
    - Can use semantic web KR but usually augments it
- **Market Segments and Players**
  - Enterprise Document Management (EDM) and search systems
  - Email autoclassifiers and inbox managers
  - Web question answering: Hakia, Powerset, TrueKnowledge, Cycorp (inCyc)...
  - Semantics for Search Result Enhancement: Yahoo! SearchMonkey
- **Some lessons with applying semantic web technology in this space**
  - Still waiting for a compelling match between technical capability and business need
    - Statistical methods are surprisingly good for basic relevance scoring (e.g., Latent Semantic Indexing, PageRank)
    - Verticals (esp. pharma) have seen some success
  - Semantic processing is only a small differentiator in these markets – you have to be great at nonsemantic queries, data import, crawling, storage, performance...

First Generation Examples: Powerset and Yahoo! SearchMonkey

- **Powerset: Natural language consumer search**
  - Web crawling, keyword indexing, relevance ranking
  - High performance for web-scale commercialization
  - Parsing of web page text with Xerox PARC's XLE system
  - Question answering with Wikipedia text and Freebase
    - Questions like "What did Microsoft acquire in 2006?" or "What did Steve Jobs say about the iPod?"
    - No standard corpora to evaluate performance
  - Acquired by Microsoft in June 2008
  - Powerset's semantic knowledge is a superset of semantic web KR

Powerset
Natural Language Search

- **Yahoo! SearchMonkey (see also Google's Rich Snippets)**
  - GreaseMonkey-style web reformatting for search
  - Yahoo's crawler indexes and interprets RDFa, microformats, delicious data, etc.
  - Display URL as an enhanced result, with standard or custom presentations
  - Incentives: "Structured data is the new SEO" (Dries Buytaert, Drupal)

SearchMonkey

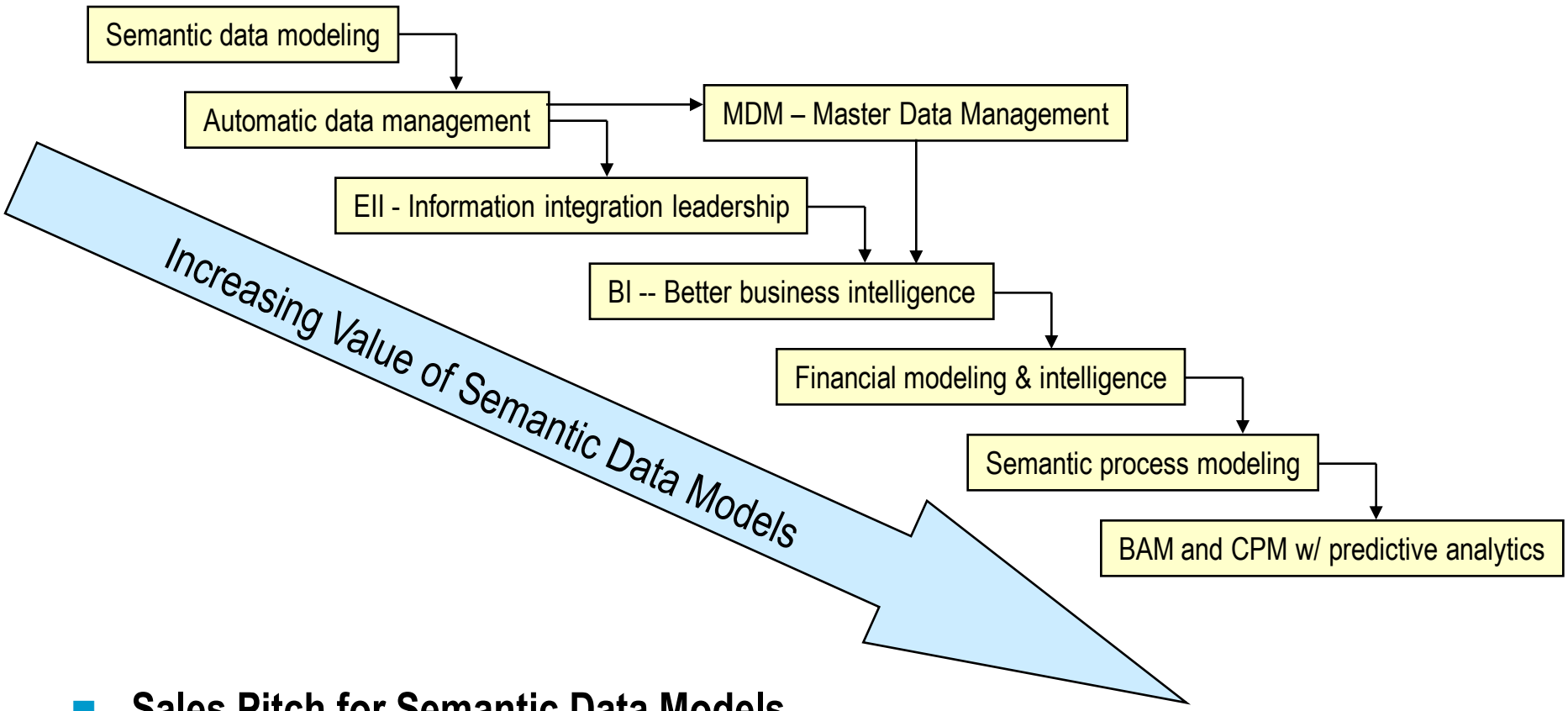
VULCAN

Second Generation Semantic Web Applications

Strategic Enterprise Information Technology

- **An only slightly newer problem type**
  - Business exploitation of structured enterprise data (RDBMS, Spreadsheet, ERP data)
    - Backwards to Data Management to reduce cost of managing, migrating, integrating
    - Forwards to Business Process Management
  - Support for unified query, analytics, and application access
    - Includes SOA integration, Enterprise Application Integration
- **Markets Segments and Players**
  - Gartner estimates that EII software and services alone is $14B/year, with 40% growth over 5 years (pre-recession numbers, though)
  - Very complex market space includes EAI, Entity Analytics, MDM, BI, BPM, CPM...
  - Huge entrenched players (IBM, SAP, Oracle...) and major consulting shops
- **Some lessons with applying semantic web technology in this space**
  - Fundamental problem is understanding the semantics from legacy systems, not in KR
  - Pure Semantic technology companies tend to be unsophisticated about the customer
  - RDF/OWL is typically too weak and must be augmented by rules, quantities, etc.
  - Raw performance is typically inferior to a well-designed RDBMS
  - Tends to be an IT sale (not Line-of-Business sale), with attendant cost pressures

Semantic Technology in Enterprise Strategic IT



- **Sales Pitch for Semantic Data Models**
  - Promote flexibility and improvisation in the face of dynamism
  - Expose business processes as rules, for governance and compliance
  - Can be driven all the way through the architecture, from SOA to CPM dashboards
- **This vision has never been proven at scale outside the lab**

VULCAN

Second Generation Non-SemWeb Example: Wolfram Alpha

- **Alpha is a computational engine, not a search engine**
  - Ultra-calculator (Mathematica) combined with a massive almanac
  - Example: "Height of Mt. Everest divided by age of youngest US President"
  - Similar to Google's special computations, but much more powerful
  - Displays the solution results using templates from Mathematica's visualization tools
- **Alpha includes 100s of manually integrated and curated data sources**
  - Topic coverage includes products, people, science, cooking, weather, travel, business, geography, music, chemistry, astronomy, physics, etc.
  - Builds on the labor of >100 people over two years, who in turn built on 20 years of Mathematica
  - NLP-driven front end with a "query-like" syntax
- **Alpha's strength is representing mathematical relations between "factual" entities found in databases**
  - These semantics are extremely deep and go beyond what is expressible in OWL

VULCAN

Third Generation Semantic Web Applications

Web 2.0 and the Socio-Semantic Web

- **A new problem type**
  - "Semantic Web should allow people to have a better online experience" – Alex Iskold, AdaptiveBlue
  - Enhance the human activities of content creation, publishing, linking my data to other data, socializing, forming community, purchasing satisfying things, browsing, etc.
  - Improve the effectiveness of advertising

- **Market Segments and Players**
  - Mashup systems and consumer-oriented semantic web services (Drupal, Ning, ...)
  - Semantic enhancements to blogs and wikis (Zemanta, Faviki, Ontoprise, Radar, ...)
  - Semantics for Social Networking (MySpace RDF service and microformats, Facebook RDF models, etc.)

- **Some lessons with applying semantic web technology in this space**
  - If we don't have semantic convergence, then semantics isn't a differentiator
  - No one really knows the design principles that allow some Web 2.0 sites to be successful and others to never get traction

VULCAN

Third Generation Example: Semantic Wikis

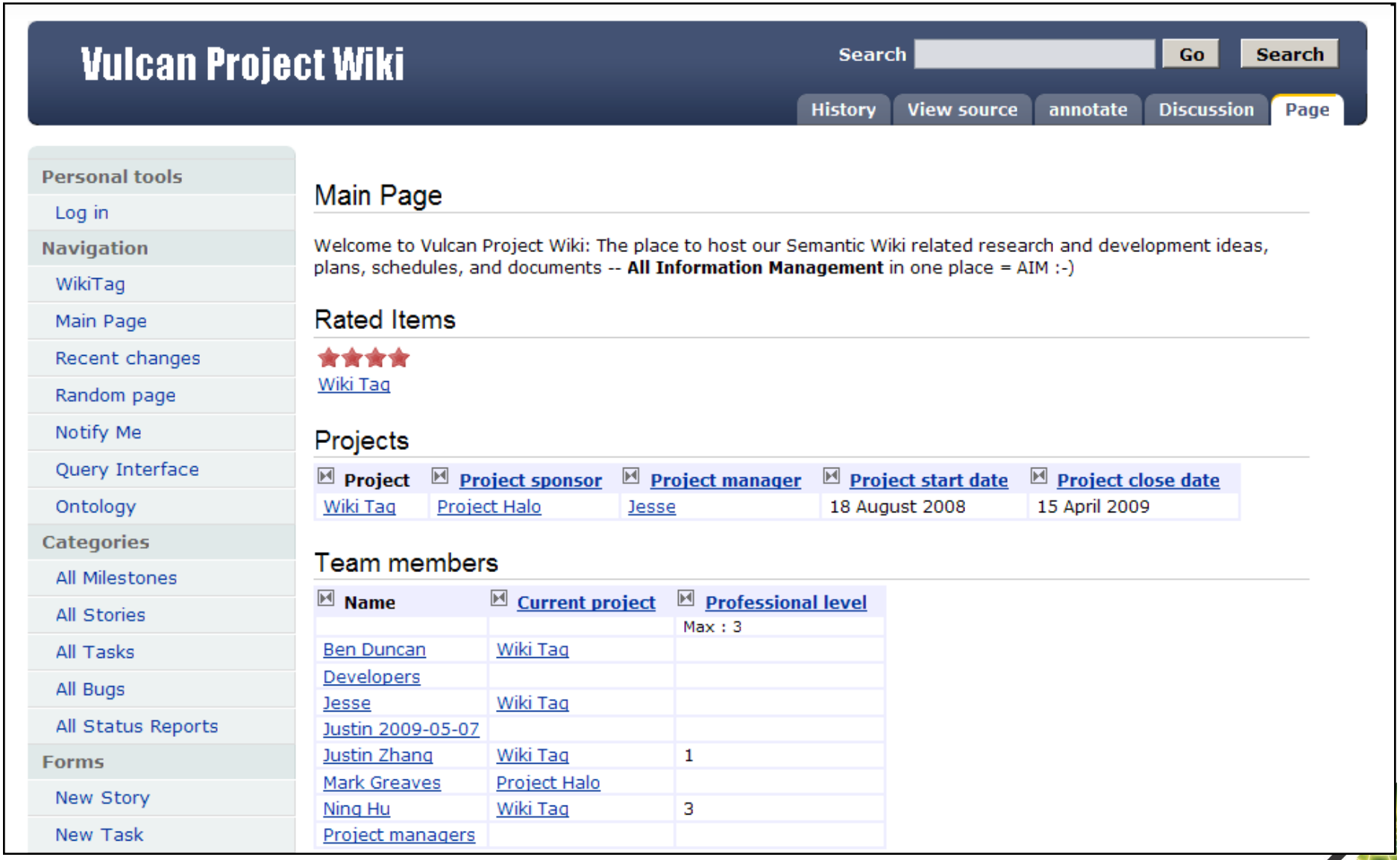
- **Wikis are tools for *Publication* and *Consensus***
- **MediaWiki (software for Wikipedia, Wikimedia, Wikibooks, etc.)**
  - Most successful Wiki software
    - High performance: 10K pages/sec served, scalability demonstrated
    - LAMP web server architecture, GPL license
  - Publication: simple distributed authoring model
    - Wikipedia: >2.9M English articles, 400K Russian, >2.5M images, #8 Alexa traffic rank
  - Consensus achieved by global editing and rollback
    - Fixpoint hypothesis, although consensus is not static
    - Gardener/admin role for contentious cases
- **Semantic Wikis apply the wiki idea to structured (typically RDFS) information**
  - Authoring includes instances, data types, vocabularies, classes
  - Natural language text used for explanations
  - Automatic list generation from structured data, basic analytics, database imports
  - Reuse of wiki knowledge
  - See e.g., http://smwforum.ontoprise.com for one powerful semantic wiki

Semantic Wiki Hypotheses:

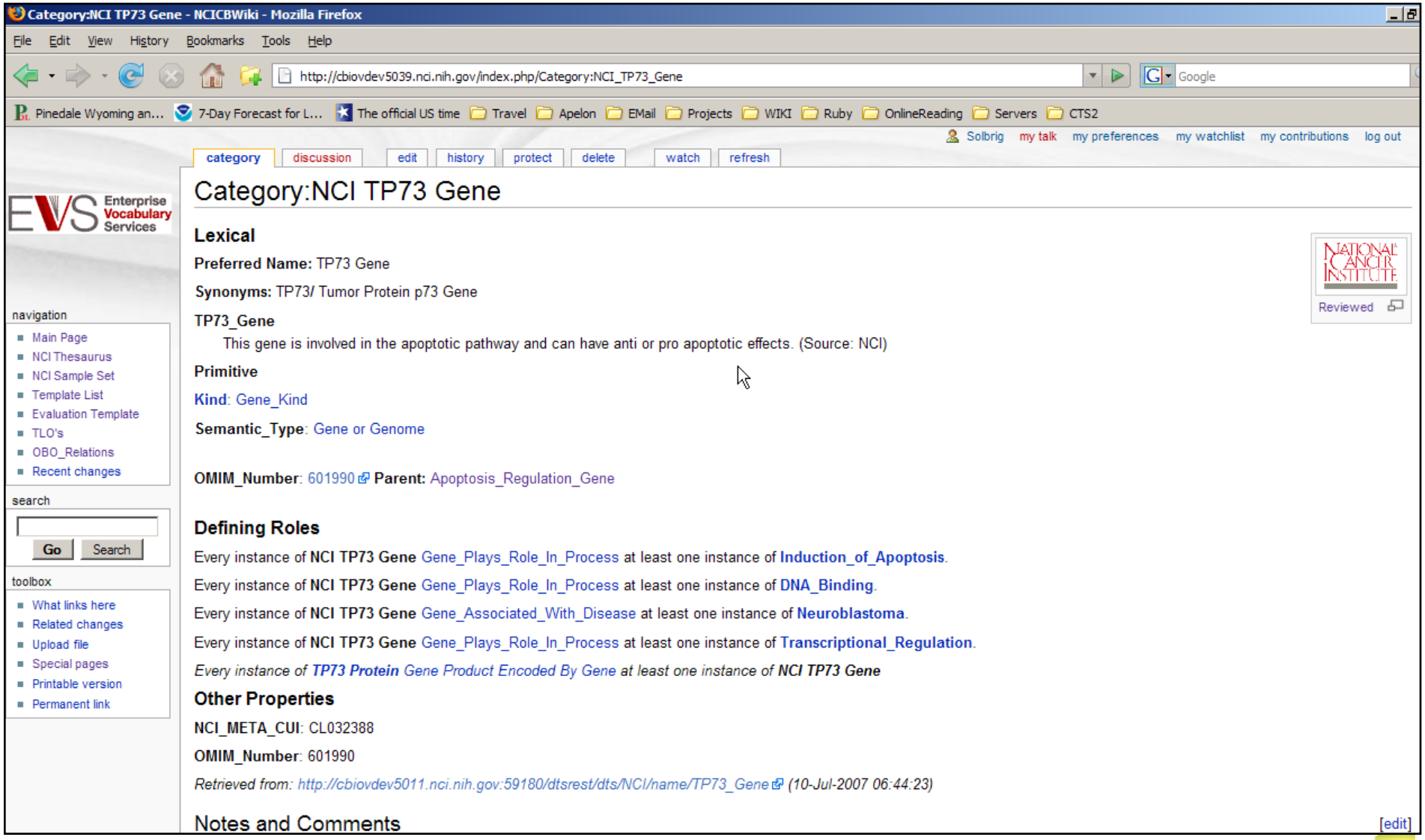
(1) Significant interesting Semantic Data can be collected cheaply

(2) Wiki mechanisms can be used to maintain consensus on vocabularies and classes

An Example of Semantic MediaWiki

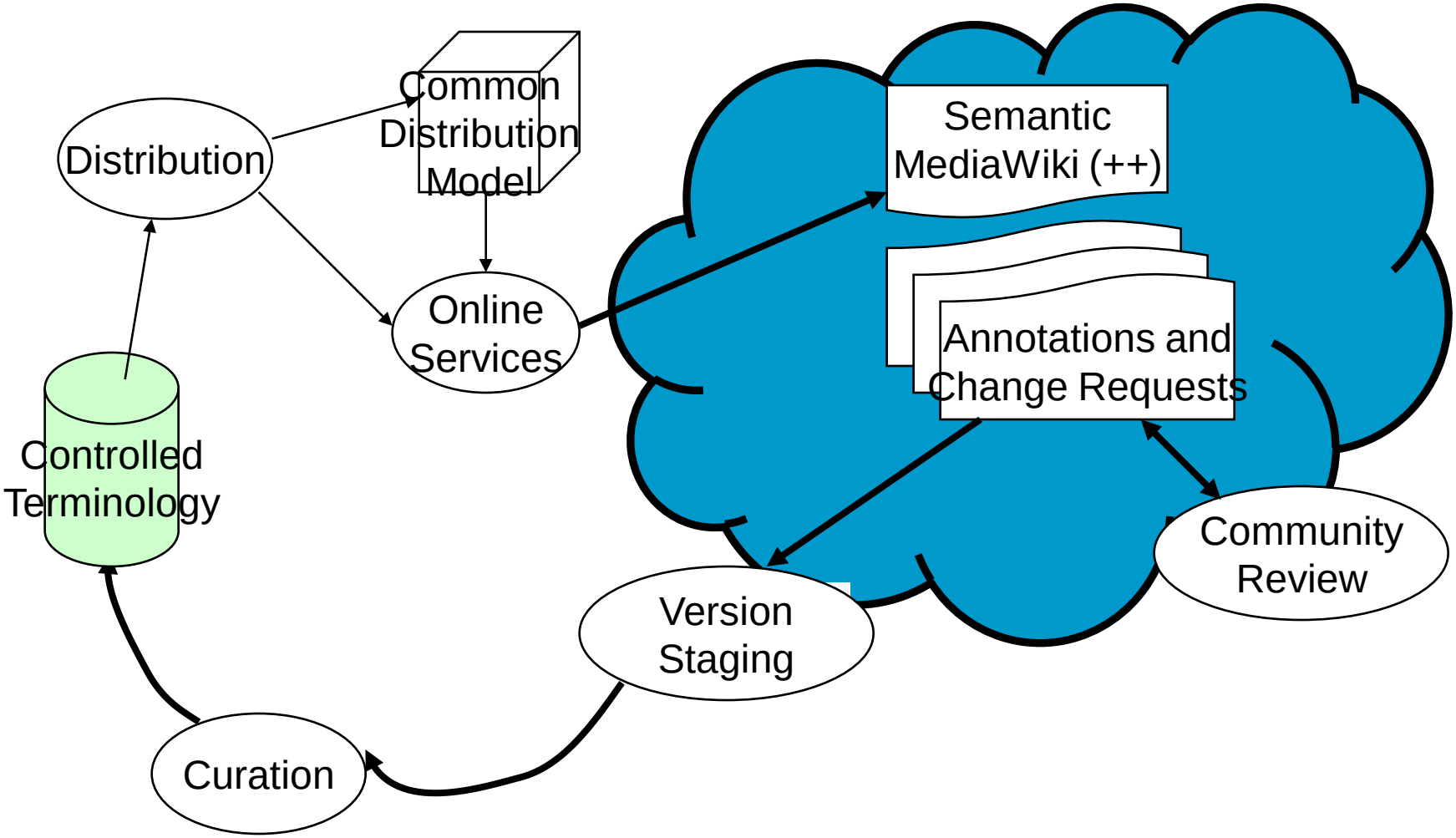


National Institutes for Health Cancer Thesaurus



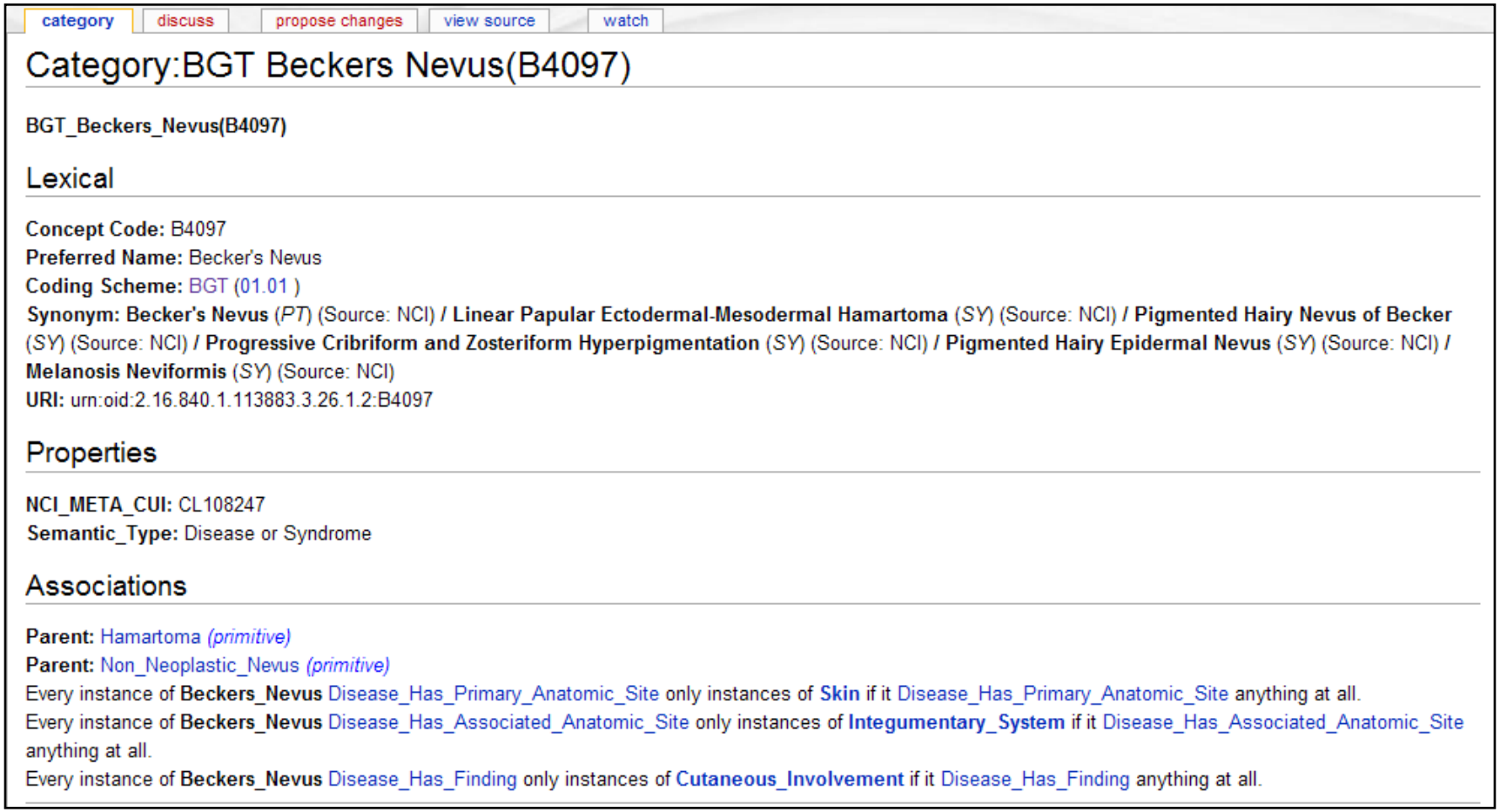
VULCAN

Real Workflows for Terminology Management



VULCAN

Example: Healthcare Vocabulary Management (HL7)



category | discuss | propose changes | view source | watch

Category:BGT Beckers Nevus(B4097)

BGT_Beckers_Nevus(B4097)

Lexical

Concept Code: B4097
Preferred Name: Becker's Nevus
Coding Scheme: BGT (01.01)
Synonym: Becker's Nevus (*PT*) (Source: NCI) / **Linear Papular Ectodermal-Mesodermal Hamartoma** (*SY*) (Source: NCI) / **Pigmented Hairy Nevus of Becker** (*SY*) (Source: NCI) / **Progressive Cribriform and Zosteriform Hyperpigmentation** (*SY*) (Source: NCI) / **Pigmented Hairy Epidermal Nevus** (*SY*) (Source: NCI) / **Melanosis Neviformis** (*SY*) (Source: NCI)
URI: urn:oid:2.16.840.1.113883.3.26.1.2:B4097

Properties

NCI_META_CUI: CL108247
Semantic_Type: Disease or Syndrome

Associations

Parent: Hamartoma *(primitive)*
Parent: Non_Neoplastic_Nevus *(primitive)*
Every instance of **Beckers_Nevus** Disease_Has_Primary_Anatomic_Site only instances of **Skin** if it Disease_Has_Primary_Anatomic_Site anything at all.
Every instance of **Beckers_Nevus** Disease_Has_Associated_Anatomic_Site only instances of **Integumentary_System** if it Disease_Has_Associated_Anatomic_Site anything at all.
Every instance of **Beckers_Nevus** Disease_Has_Finding only instances of **Cutaneous_Involvement** if it Disease_Has_Finding anything at all.

- Large-scale terminology management at www.biomedgt.org.
- Semantic MediaWiki+ blends a database and a wiki

VULCAN

Semantic Data in HL7 (www.BioMedGT.org)

Facts about BGT Beckers Nevus(B4097) RDF feed

BGT Disease Has Associated Anatomic Site(R100)	BGT Integumentary System(B13080)
BGT Disease Has Finding(R108)	BGT Cutaneous Involvement(B39743)
BGT Disease Has Primary Anatomic Site(R101)	BGT Skin(B12643)
BGT NCI META CUI(P208)	CL108247
BGT Semantic Type(P106)	Disease or Syndrome
BGT Synonym(P104)	Becker's Nevus, Linear Papular Ectodermal-Mesodermal Hamartoma, Pigmented Hairy Nevus of Becker, Progressive Cribriform and Zosteriform Hyperpigmentation, Pigmented Hairy Epidermal Nevus, and Melanosis Neviformis
DCTerms hasVersion(hasVersion)	BGT/Versions/01.01
Has default form	LexWiki BGT Form
LexWiki Concept Code	B4097
LexWiki Preferred Name	Becker's Nevus
LexWiki URI	urn:oid:2.16.840.1.113883.3.26.1.2:B4097
SKOS inScheme(inScheme)	BGT

This category currently contains no articles or media.

Categories: BGT Hamartoma(B3248) | BGT Non Neoplastic Nevus(B4110)

- Subject-matter experts give simple, authorable statements
- Adds Protégé-managed rules for terminological coherence

VULCAN

Commercial Semantic Wikis: Chickipedia

chickipedia

search GO

Categories | All Chicks | Random Chick | Help | Log In / Create Account

Add New Chick

Embed

Heidi Klum

Article | Photos | Videos | The Buzz

Add to My Babes | Talk | View source | History | Links In

"Going blonde is like buying yourself a light bulb!"

Birthday: 6/1/1973
Nickname: The Klum
Birthname: Heidi Klum
Hookups: Seal, Flavio Briatore
Hometown: Bergisch Gladbach, North Rhine-Westphalia, West Germany
Assets:
Vices:

Sign: Gemini
Height: 5'9
Job: Supermodel, TV host
Hobbies: Modeling, Talk shows
Ethnicity: White
Country of Origin: Germany

36
27
36

Similar Chicks

Chicks She's Worked With

Adriana Lima, Gisele Bündchen, Alessandra Ambrosio, Karolina Kurkova, Niki Taylor, Marisa Miller, Oprah Winfrey, Halle Berry, Charlize Theron, Anne Hathaway, Gisele Bündchen, Alessandra Ambrosio, Karolina Kurkova, Niki Taylor, Marisa Miller, Oprah Winfrey

Introduction

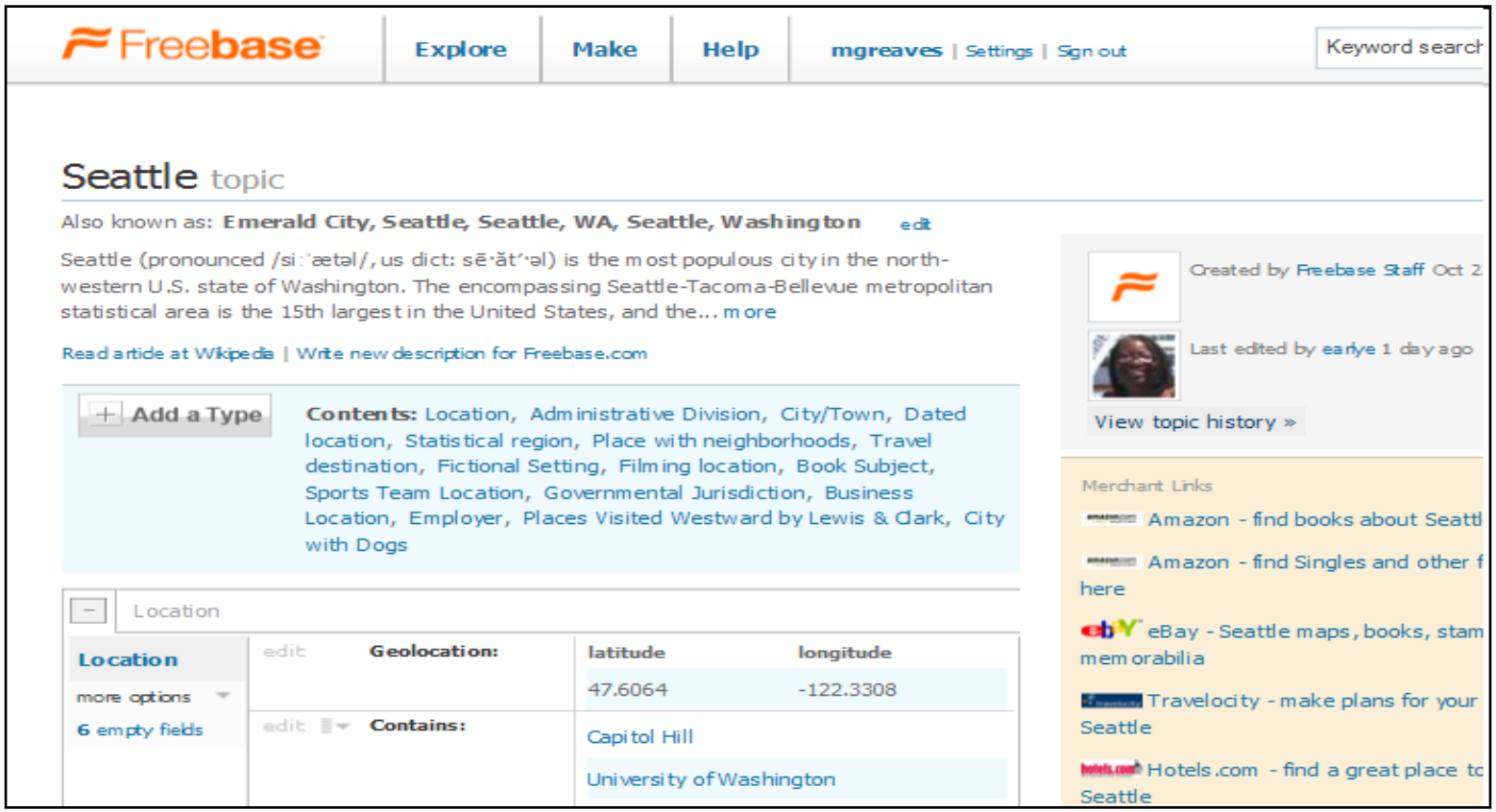
Along with pretzels, BMWs and Kraftwerk, Heidi Klum is one of our favorite German exports. And of the four, she's the one we'd most like to spend a long weekend with. Or at least she's right behind Kraftwerk. Voted #25 in the AskMen.com's Top 99 Women 2008 Edition.

Life Story

Raised by her parents Gunther and Erna in Bergisch Gladbach, a small city outside Cologne, Heidi entered a modeling contest at age 18, and out of 25,000 contestants, she was voted the winner, which was announced on the German talk show Gottschalk Late Night, a German comedy show in which no laughter is allowed. Buoyed by a modeling contract worth $300,000, Heidi decided to

VULCAN

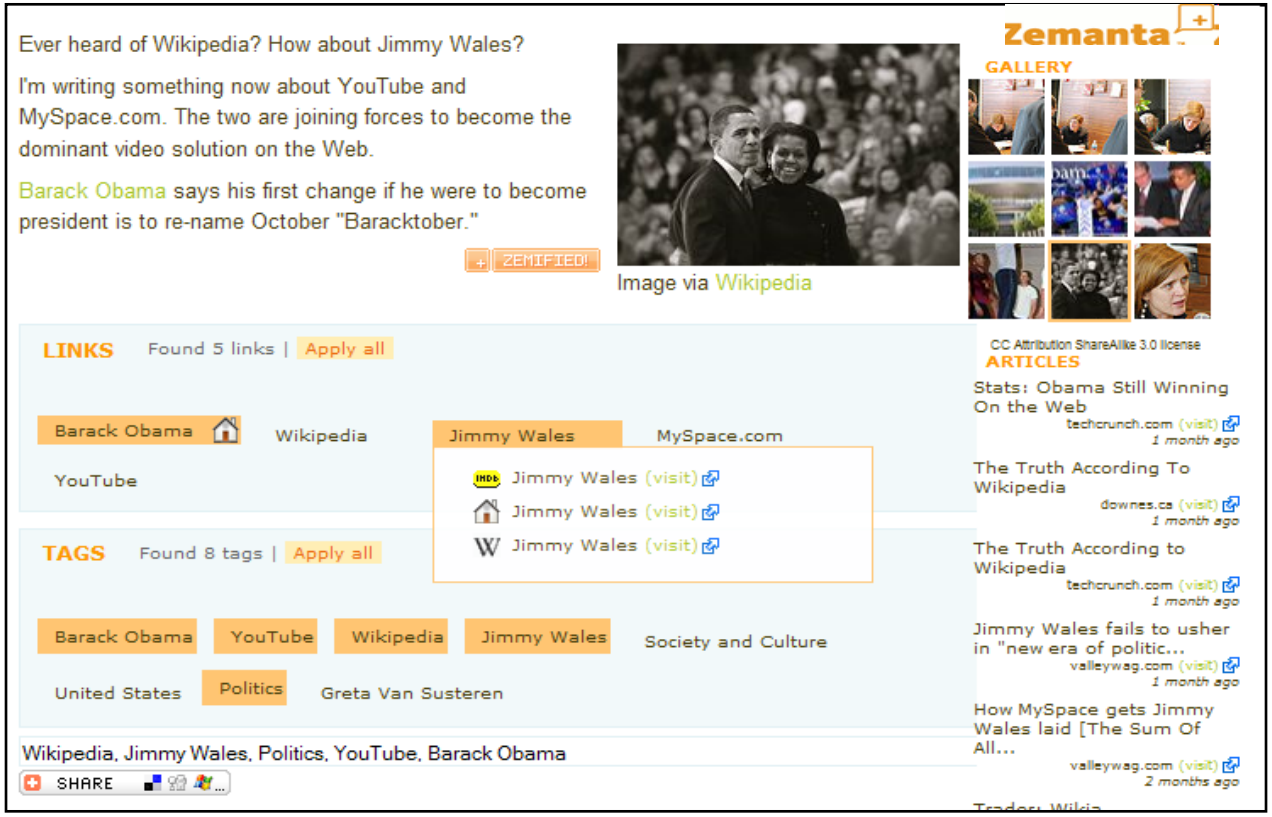
Web Knowledge Bases: Metaweb and Freebase



- Massive amounts of almanac-style RDF data (Creative Commons license) that is readily available from partners
- Social authoring tools and wiki-style consensus combined with controlled reconciliation by Metaweb personnel
- Data outsourcing model for long-tail startups

VULCAN

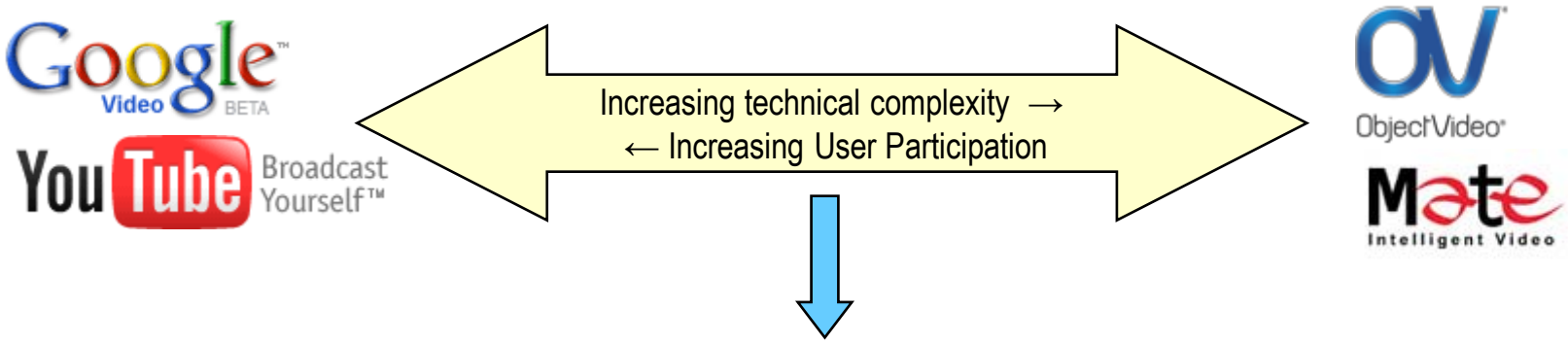
Semantic Blogger Support: Zemanta



- **Automatic link, image, keyword, tag suggestions for bloggers (and email)**
  - Average semi-professional blogger spends ~20 mins adding "decorative" content
- **Accuracy is guaranteed because users explicitly add the suggestions**
  - Zemanta inserts RDFa and standard semantic markup in the background
  - Includes user specified friends/feeds/photos/etc as well as standard ones

VULCAN

Semantic Video Wikis: US Football Video



- Social tag-based characterization
- Keyword search over tag data
- Inconsistent semantics
- Easy to engineer

Semantic Entertainment (SMW 1.3)

- Social database-style characterization
- Category/Property+ wiki text query
- Semantic consistency via wiki mechanisms
- Easy to engineer

- Algorithm-based object characterization
- Database-style search
- Consistent semantics
- Extremely difficult to engineer

VULCAN

navigation | create a new play | views | toolbox

- Log in / create account

search

Go Search

browse

- Seahawks Games
- SEA-GB
- SEA-STL
- 49ers-SEA
- SEA-EAGLES
- All plays for all games

Main Page

Welcome to *Semantic Football*

This is a sandbox site for putting semantic markup information on football video clips.

This site is currently restricted to authorized users only. Please contact related personnels to obtain a user account if you don't have one yet.

Contents [hide]

Queries to demonstrate semantics

Some Interesting Queries

A list of **Interesting Queries** that we can do now or make it happen easily with some extra work.

Wikipedia NFL/Seahawks Player Data Import

We can import player's data (birthdate, college, team, etc.) from Wikipedia/DbPedia or Freebase, so that contents available from external sources does NOT need be recreated again, and we can even sync with the external data in some way.

All Seahawks Players

All Plays in Seahawks Games

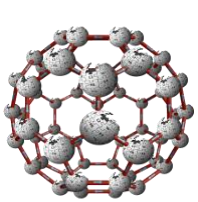
Currently, there are 639 plays in the Wik.

where,

- total of 152 plays for Seattle vs Green Bay game
- total of 160 plays for Seattle vs St. Louis game
- total of 163 plays for San Francisco vs Seattle game
- total of 162 plays for Seattle vs Philadelphia game

Talk Outline: The Maturing Semantic Web

- **The Origins of the Semantic Web**
  - DARPA's DAML Program
  - RDF, OWL, and the Semweb Infrastructure
- **Semantic Web Evolution to 2009**
  - Three Generations of Semantic Dreams
  - Markets and Companies
- **The Fourth Generation**
  - A Scalable Revolution



Fourth Generation Semantic Web

The Web of Data meets the Future Internet

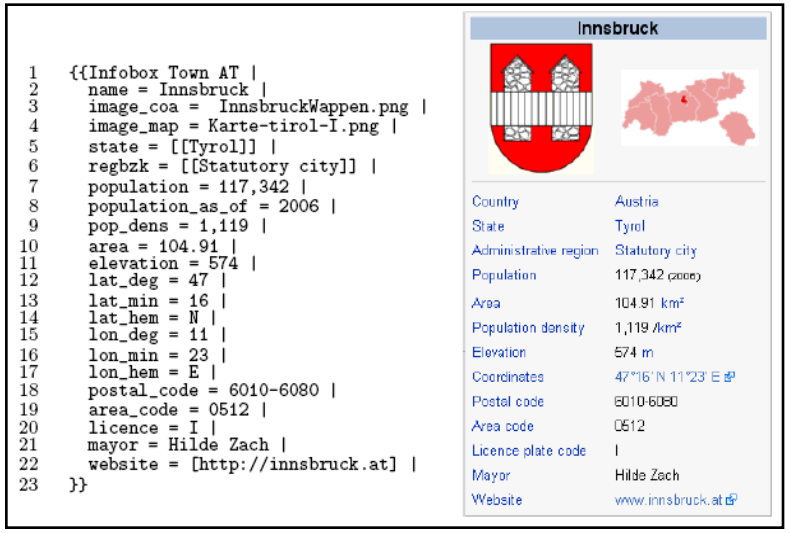
- **A problem of scale**
  - The number of Internet devices is starting to explode (again!)
    - Mobile devices, embedded systems, and sensors
    - In 2008, Google reported 1 trillion unique URLs, ~200M web sites
    - Total 2008 web page estimates are ~30 billion (significant variation in these estimates)
  - Gartner (May 2007, Report G00148725)
    - "By 2012, 70% of public Web pages will have some level of semantic markup, 20% will use more extensive Semantic Web-based ontologies"
  - Can Semantic Web technologies work at web scales?
    - Sindice (www.sindice.com) is now indexing >10B triples/microformats over 100M pages
    - 20% of 30 billion pages @ 1000 triples per page = **6 trillion triples**
    - 30 billion and 1000 are underestimates

Material from Frank van Harmelen, Vrije Universiteit, Amsterdam

- **Some lessons with applying semantic web technology in this space**
  - Does the Semantic Web have the Google Property?
  - Can we exploit billions of triples, microformats, ontologies, rules, and services?
    - Are Semantic Web systems deployable on parallel web architectures, friendly to out-of-core algorithms, and compatible with giant databases?
  - Is there a scaling limit to useful, profitable Semantic Web implementations?

Fourth Generation Example: DBpedia

- **Mine Wikipedia for assertions**
  - Mainly from Wikipedia Factboxes
    - ~23M triples
  - Category assertions
- **DBpedia 3.2 dataset (Oct 08 Wikipedia)**
  - ~2.6M things, ~274M triples
    - 213K persons, 328K places, 57K music albums, 36K films, 609K links to images, 3.2M links to relevant external web pages, 4.9M links into RDF datasets
  - Classifications via Wikipedia categories, YAGO, and WordNet synsets
  - One of the largest broad knowledge bases in the world
- **V.3.3 released July 3 (May 09 Wikipedia)**
- **Simple queries over extracted data**
  - "Things near the Eiffel Tower"
  - "The official websites of companies with more than 50000 employees"
  - "Soccer players from team with stadium with >40000 seats, who were born in a country with more than 10M inhabitants"



```
{{Infobox Town AT |
  name = Innsbruck |
  image_coa =  InnsbruckWappen.png |
  image_map = Karte-tirol-I.png |
  state = [[Tyrol]] |
  regbzk = [[Statutory city]] |
  population = 117,342 |
  population_as_of = 2006 |
  pop_dens = 1,119 |
  area = 104.91 |
  elevation = 574 |
  lat_deg = 47 |
  lat_min = 16 |
  lat_hem = N |
  lon_deg = 11 |
  lon_min = 23 |
  lon_hem = E |
  postal_code = 6010-6080 |
  area_code = 0512 |
  licence = I |
  mayor = Hilde Zach |
  website = [http://innsbruck.at] |
}}
```


UNIVERSITÄT LEIPZIG

DBpedia

Freie Universität Berlin

OPENLINK SOFTWARE

VULCAN

DBpedia for Users

Query Wikipedia like a database

UNIVERSITÄT LEIPZIG pedia

Query Wikipedia

This semantic database contains over 10 million statements extracted from the English Wikipedia.

search for queries | Most popular | Upcoming

- Tennis players from Moscow
- Sitcoms set in NYC
- Soccer player with tricot nr. 11, playing for a club having a stadium with >40.000 seats, born in a country with >10M inhabitants
- People influenced by Friedrich Nietzsche
- Films longer than 5 hours
- Space Missions
- Film music composer born 1965
- People being 1.80m tall
- List of Web browser software
- Mayors of US cities higher than 1000m
- Pictures of American guitarists
- Battles in Saxony
- What connects Innsbruck and Leipzig
- Hip hop CDs from Texas Artists
- Scientists and their doctoral advisors

<< 1 >>

Soccer player with tricot nr. 11, playing for a club having a stadium with >40.000 seats, born in a country with >10M inhabitants

Subject	Predicate	Object	
?player	currentclub	?club	[-]
?player	clubnumber	11	[-]
?player	countryofbirth	?country	[-]
?club	capacity	>40000	[-]
?country		>10000000	[-]

[+]

GDP_PPP (11)
population_estimate (10)
population_census (9)
established_date2 (8)
established_date1 (6)
established_date3 (5)
GDP_nominal (3)
accessionEUdate (2)

Click on a column head… this page. Results: 10

10 results found in 0.0…

Nr.	?player		?country	>40000	>10000000
1	Cicinho	Real Madrid	Brazil	80354	187560000
2	Gonzalo Fierro	Colo-Colo	Chile	62000	16432674
3	Lukas Podolski	FC Bayern Munich	Poland	69901	38536869
4	Mark González	Liverpool F.C.	South Africa	45362	47432000
5	Michael Thurk	Eintracht Frankfurt	Germany	52000	82438000
6	Ramón Morales	Chivas de Guadalajara	Mexico	72480	107784179
7	Robin van Persie	Arsenal F.C.	Netherlands	60432	16336346
8	Stefano Mauri	S.S. Lazio	Italy	82656	58751711

DBpedia Mobile

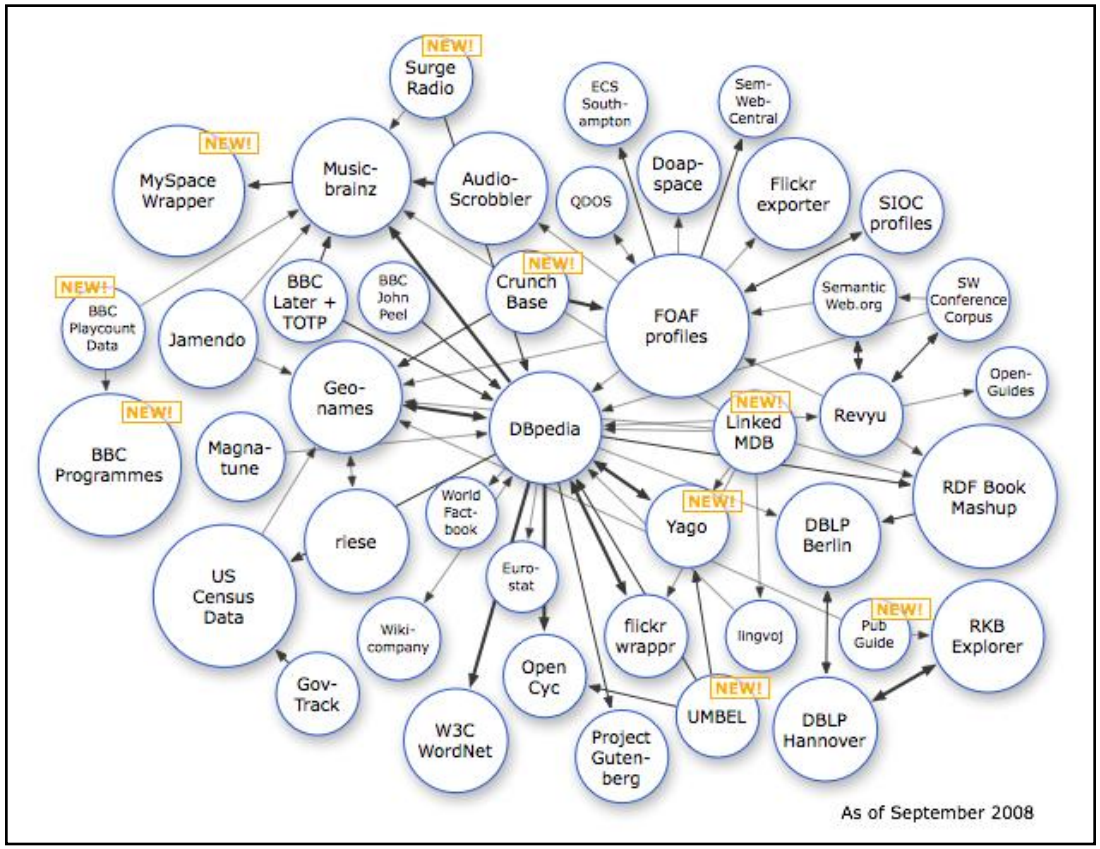
VULCAN

Fourth Generation Example: Linking Open Data

- **Goals**
  - Create a single, simple set of rules for publishing and linking RDF data
  - Build a data commons by making open data sources available on the Web as RDF
  - Set RDF links between data items from different data sources

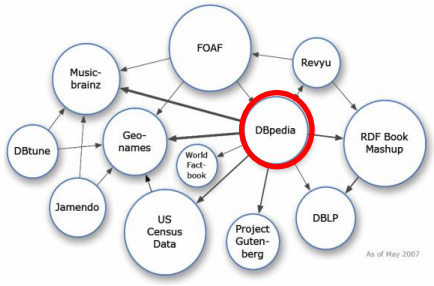
- **May 2009 LOD dataset**
  - ~4.7B triples, and ~140M RDF interlinks, and growing faster than I can track
  - Database linkage means that LOD will soon be impossible to count except by order of magnitude

LINKINGOPENDATA
W3C SWEO Community Project

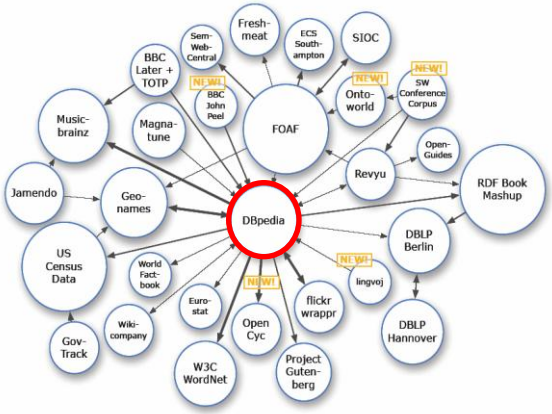


VULCAN

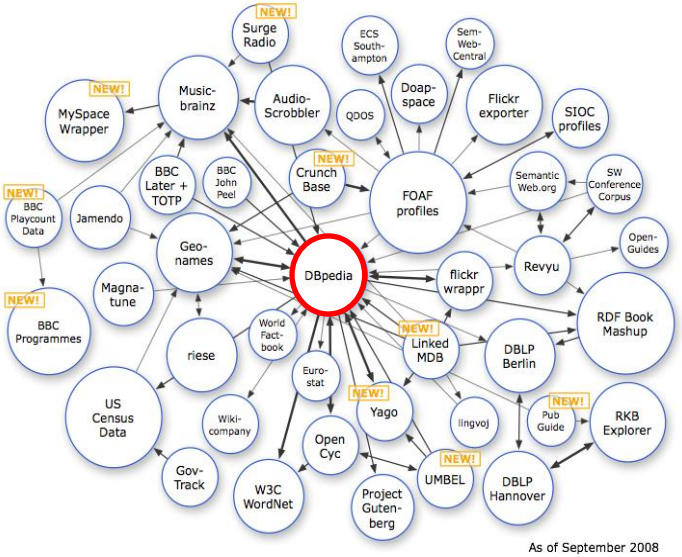
The Growing Web of Data



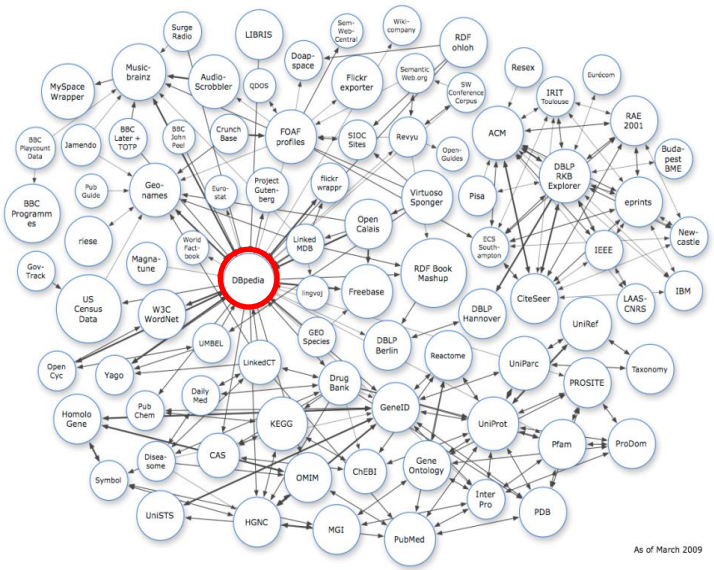
May 2007



September 2007



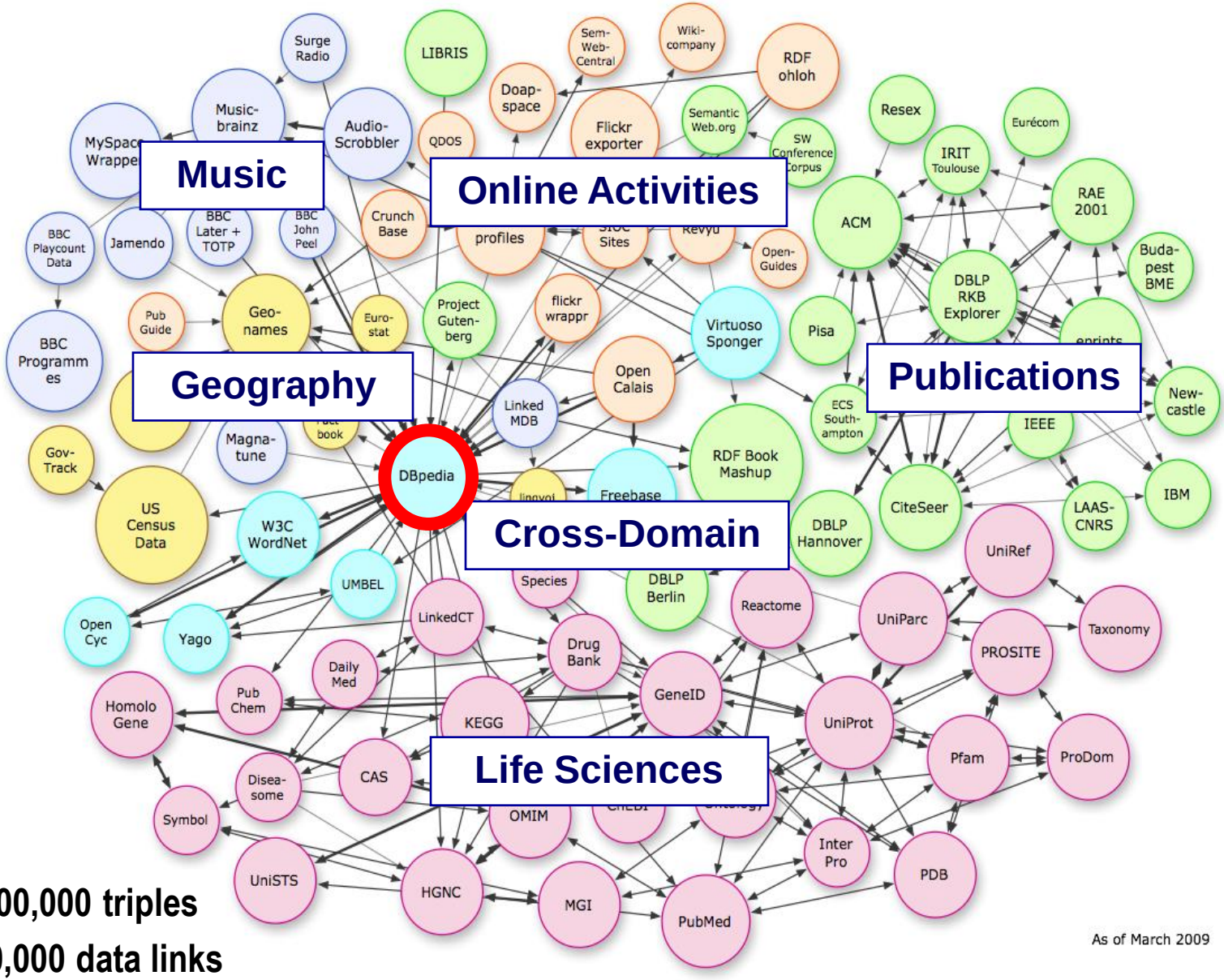
September 2008



March 2009

VULCAN

Topic Distribution in the Linked Datasets



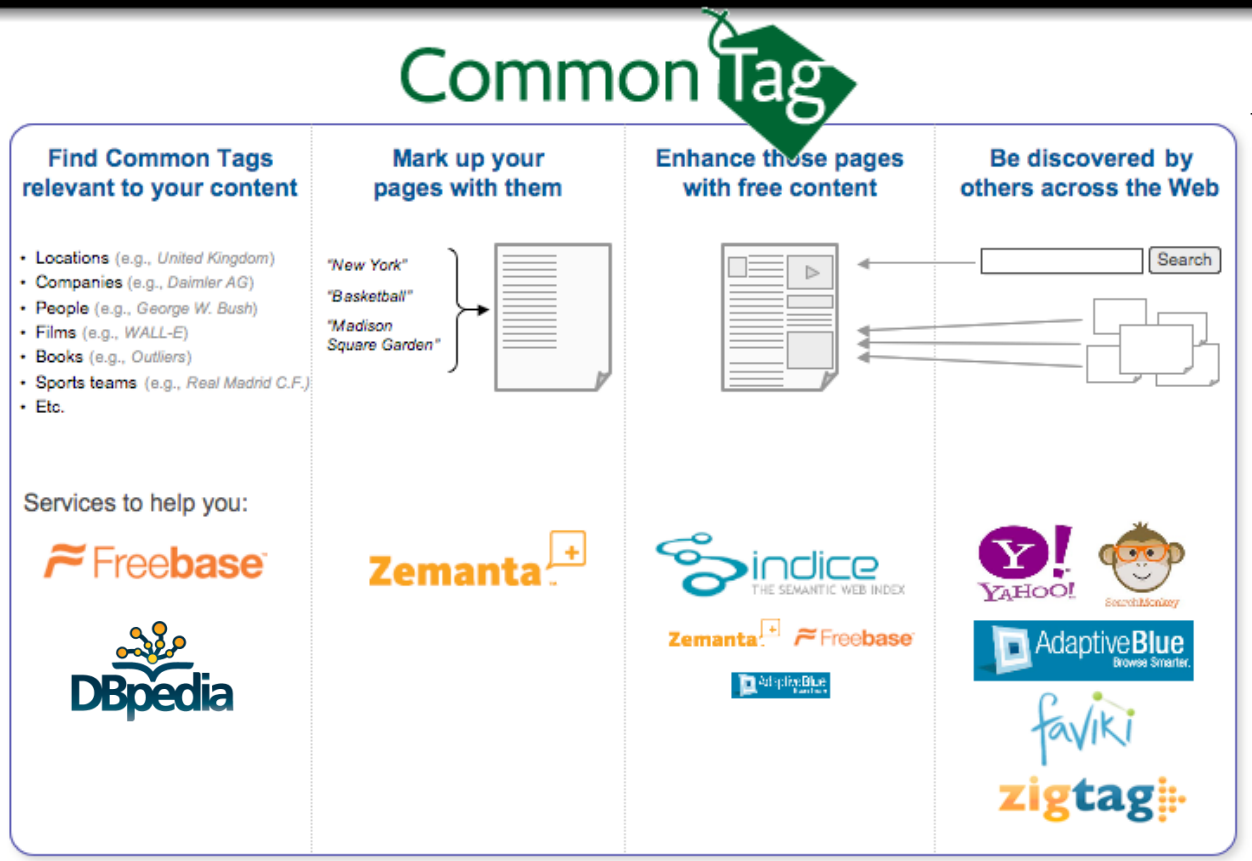
4,500,000,000 triples

180,000,000 data links

As of March 2009

VULCAN

Common Tag Specification



- **Instead of tagging with language terms, tag with terms + RDFa**
  - Distinguish between "jaguar" the animal, the car company, and the operating system
  - Provides metadata for each Common Tag and relations to other Common Tags
    - The Barack Obama Common Tag includes <employment, President of the United States> and <spouse, Michelle Obama>
- **More discoverable, more connected, more web-like, more useful**

VULCAN

Semantic Dynamism at Web Scale

- **Semantics are always changing**
  - Per minute, there are:
    - 100 edits in Wikipedia (144K/day)
    - 200 tags in del.icio.us (288K/day)
    - 270 image uploads to flickr (388K/day)
    - 1100 blog entries (1.6M/day)
  - Will the Semantic Web be less dynamic?

- **There is no "right ontology"**
  - Ontologies are abstractions
    - Different applications lead to different ontologies
    - Ontology authors make design choices all the time
  - Google Base: >250K schemas
  - "Ontologies = Politics"

- **Intentionally false material (Spam)**
  - Lesson of the HTML <META> tag

Material from Denny Vrandečić, AIFB

How Do We Use this Dynamic Data for Decision Support?

VULCAN

Fourth Generation Application: The Large Knowledge Collider

- **EC Framework 7 Program**
  - Lead partners: Univ. Innsbruck and Vrije University Amsterdam, plus 12 partners
- **Goals of LarKC – Scaling to Infinity**
  - A platform for massive distributed incomplete reasoning
  - Remove the scalability barriers of currently existing reasoning systems for the Semantic Web.
  - Combine reasoning/retrieval and search
  - Want to trade off answer quality and answer timeliness
- **Reasoning pipeline**
  - Heavy emphasis on probability, decision theory, anytime algorithms
  - Plugin architecture, with sampling
  - Explicit cost models
- **Public releases of LarKC platform, with APIs**
- **Encourage participation through Thinking@home**
  - Kind of like SETI@Home

Fourth Generation Application: The Large Knowledge Collider

- EC Framework 7 Program
  - Lead partners: Univ. Innsbruck and Vrije University Amsterdam, plus 12 partners
- Reasoning pipeline
  - Heavy emphasis on probability, decision theory, anytime algorithms
- **Exploiting *web-scale semantics* is the new frontier**
  - Generations 1 and 2 used web resources to support classical KR approaches
  - Generation 3 (social semantic web) leverages web social patterns for KR
- **Fourth generation applications address general *web-scale* KR**
  - currently existing reasoning systems for the Semantic Web.
  - Combine reasoning/retrieval and search
  - Want to trade off answer quality and answer timeliness
- Encourage participation through Thinking@home
  - Kind of like SETI@Home

VULCAN

Fourth Generation Application: The Large Knowledge Collider

- EC Framework 7 Program
  - Lead partners: Univ. Innsbruck and Vrije
- Reasoning pipeline
  - Heavy emphasis on probability,
- **The real money in semantics will be made in apps/tools that exploit web-scale data**
  - The cost of semantic data creation is going to zero
  - The size of semantic data is going to web-scale
- ***If LarKC is successful, this could be as big as PageRank™!***

the Semantic Web.

- Combine reasoning/retrieval and search
- Want to trade off answer quality and answer timeliness

- Encourage participation through Thinking@home
  - Kind of like SETI@Home

VULCAN

Summing up: The Maturing Semantic Web

- **In mid-2004...**
  - RDF and OWL had just been standardized
  - Advances were made via traditional corporate/public R&D programs
  - The first wave of semantic web startups (many of which have since failed)
  - US Government implementations were technically very sophisticated, but fully custom and had no web involvement
  - A few early conferences (ISWC, SemTech) and session tracks

Summing up: The Maturing Semantic Web

- **In mid-2004...**
  - RDF and OWL had just been standardized
  - Advances were made via traditional corporate/public R&D programs
  - The first wave of semantic web startups (many of which have since failed)
  - US Defense implementations were technically very sophisticated, but fully custom and had no web involvement
  - A few early conferences (ISWC, SemTech) and session tracks
- **Now in 2009...**
  - The Semantic Web is the most exciting thing happening on the web
  - RDF assertions scaling into the billions, with little to no programmatic control
  - Search majors are starting to develop products
  - Bestbuy is publishing store descriptions and hours in RDFa

Summing up: The Maturing Semantic Web

- **In mid-2004...**
  - RDF and OWL had just been standardized
  - Advances were made via traditional corporate/public R&D programs
  - The first wave of semantic web startups (many of which have since failed)
  - US Defense implementations were technically very sophisticated, but fully custom and had no web involvement
  - A few early conferences (ISWC, SemTech) and session tracks
- **Now in 2009...**
  - The Semantic Web is the most exciting thing happening on the web
  - RDF assertions scaling into the billions, with little to no programmatic control
  - Search engines are starting to develop products
  - Bestbuy is publishing store descriptions and hours in RDFa
- **I've Got that Scaling Feeling (i.e., the Google Property)**
  - Incentives are starting to falling into place with the search engines
  - The Linked Data cloud is getting bigger and better
  - The infrastructure to build non-tools-based companies is in place
  - Conceptual Graph Results are directly relevant

VULCAN

Спасибо

(Thank You)

Disclaimer: The preceding slides represent the views of the author only.
All brands, logos and products are trademarks or registered trademarks of their respective companies.

VULCAN