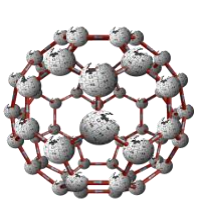
The Evolving Semantic Web: From Military Technology to Venture Capital

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

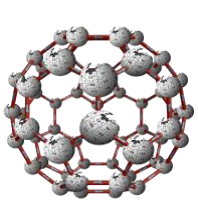
Talk Outline: The Evolving Semantic Web

- **US Semantic Web R&D**
  - DARPA's DAML Program
- **Semantic Web Evolution to 2008**
  - Three Generations of Semantic Dreams
  - Markets and Companies
- **The Fourth Generation**
  - A Scalable Revolution



Talk Outline: The Evolving Semantic Web

- **US Semantic Web R&D**
  - DARPA's DAML Program
- **Semantic Web Evolution to 2008**
  - Three Generations of Semantic Dreams
  - Markets and Companies
- **The Fourth Generation**
  - A Scalable Revolution



Preliminaries on US Computer Science Funding

- **Major US Funding Organizations for Information Technology R&D**
  - US National Science Foundation – mainly individual professors
  - Department of Energy Office of Science – large-scale scientific computing
  - Department of Defense – DARPA, AFRL, ARL, ONR

- **DARPA = Defense Advanced Research Projects Agency**
  - Long Range R&D Organization of the US Department of Defense
    - Established 1958 as a US response to the Soviet launch of Sputnik
    - Pursues high-risk, high-payoff basic and applied research with military application
  - Chartered to Prevent Technological Surprise
  - Recent budgets are $3-3.5B/year across all efforts and scientific disciplines

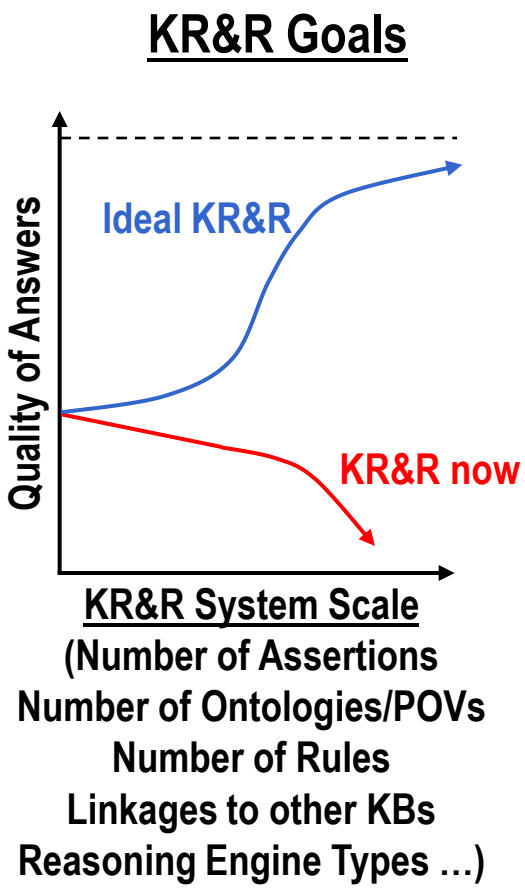
- **DARPA is the major source for large-scale AI research funding in the US**

- **Lightweight organizational model**
  - "120 Program Managers with a common travel agent"
  - No dedicated facilities beyond simple office space
  - Program Managers have a greater degree of operational control than EU Project Officers
  - http://www.darpa.mil has current programs, solicitations, lists, areas of interest

VULCAN

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 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
- Use a 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)

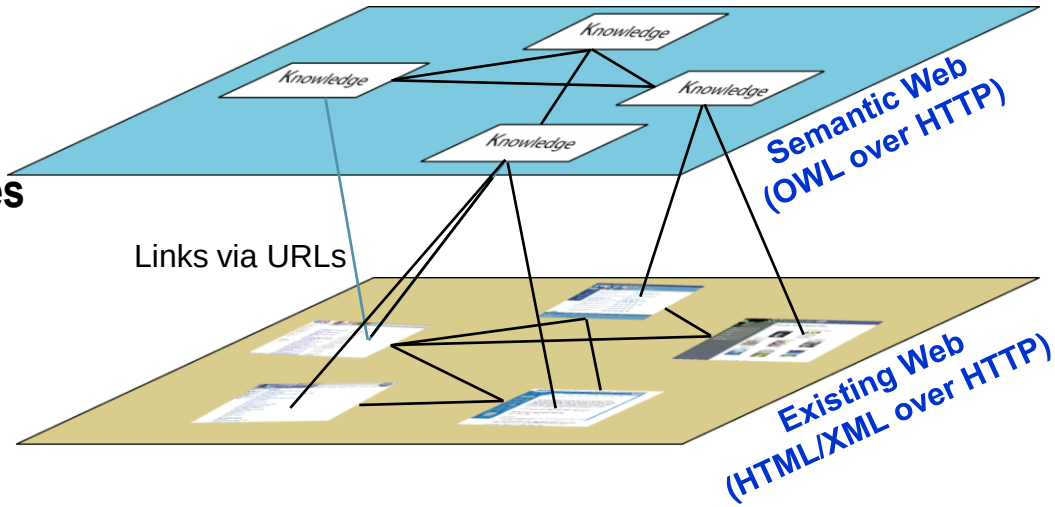
Build definition and markup tools

Link new knowledge to existing web page elements

Test design approach with operational pilots in the US DoD

Partner with parallel EU efforts to standardize the new web languages

Computers require explicit knowledge to reason with web pages



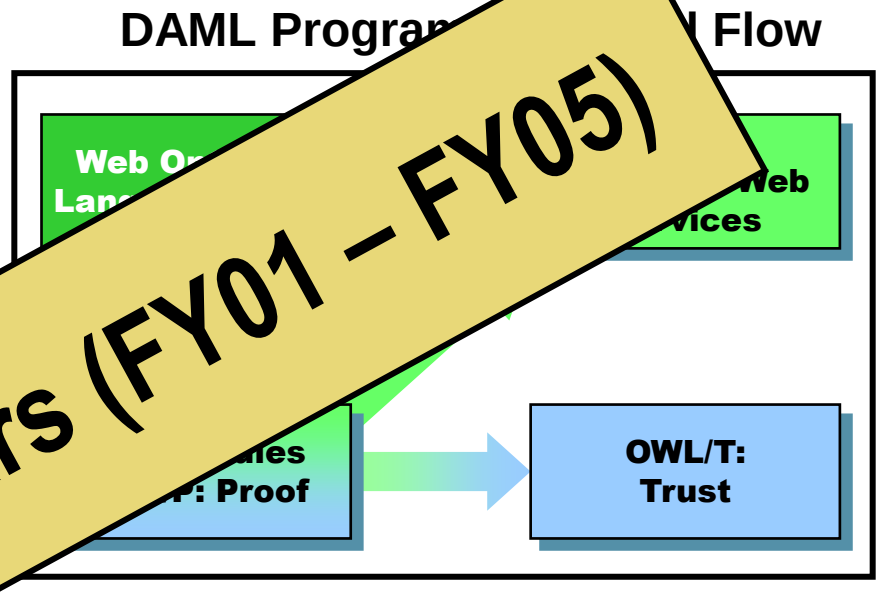
People use implicit knowledge to reason with web pages

VULCAN

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 DoD pilots and prototypes**

~$45M over 5 years (FY01 – FY05)



Completed standards process

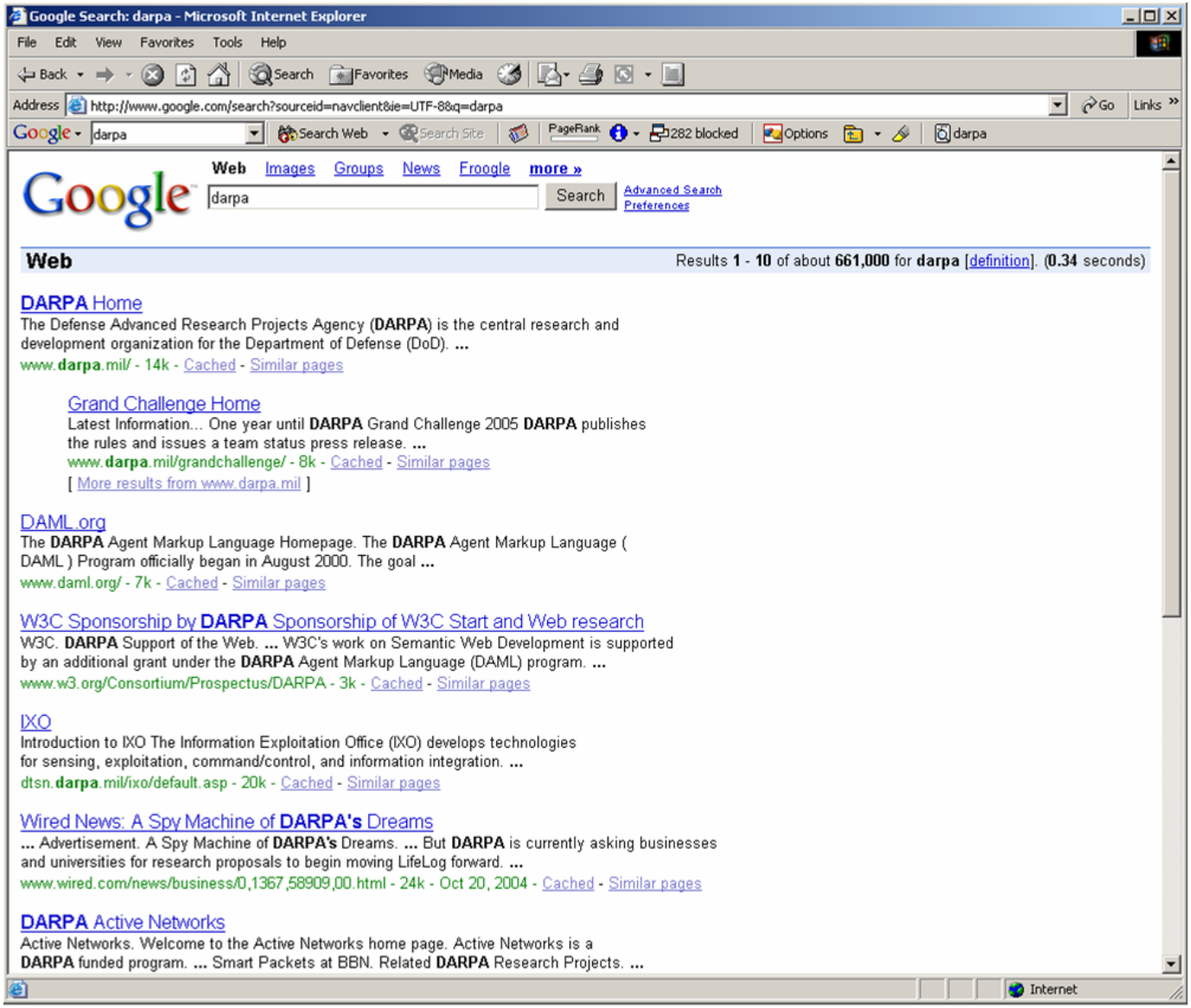
Started standards process

Unfinished

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

VULCAN

Another Measure of Impact



#2

#3

Google "darpa" on 10/21/04

VULCAN

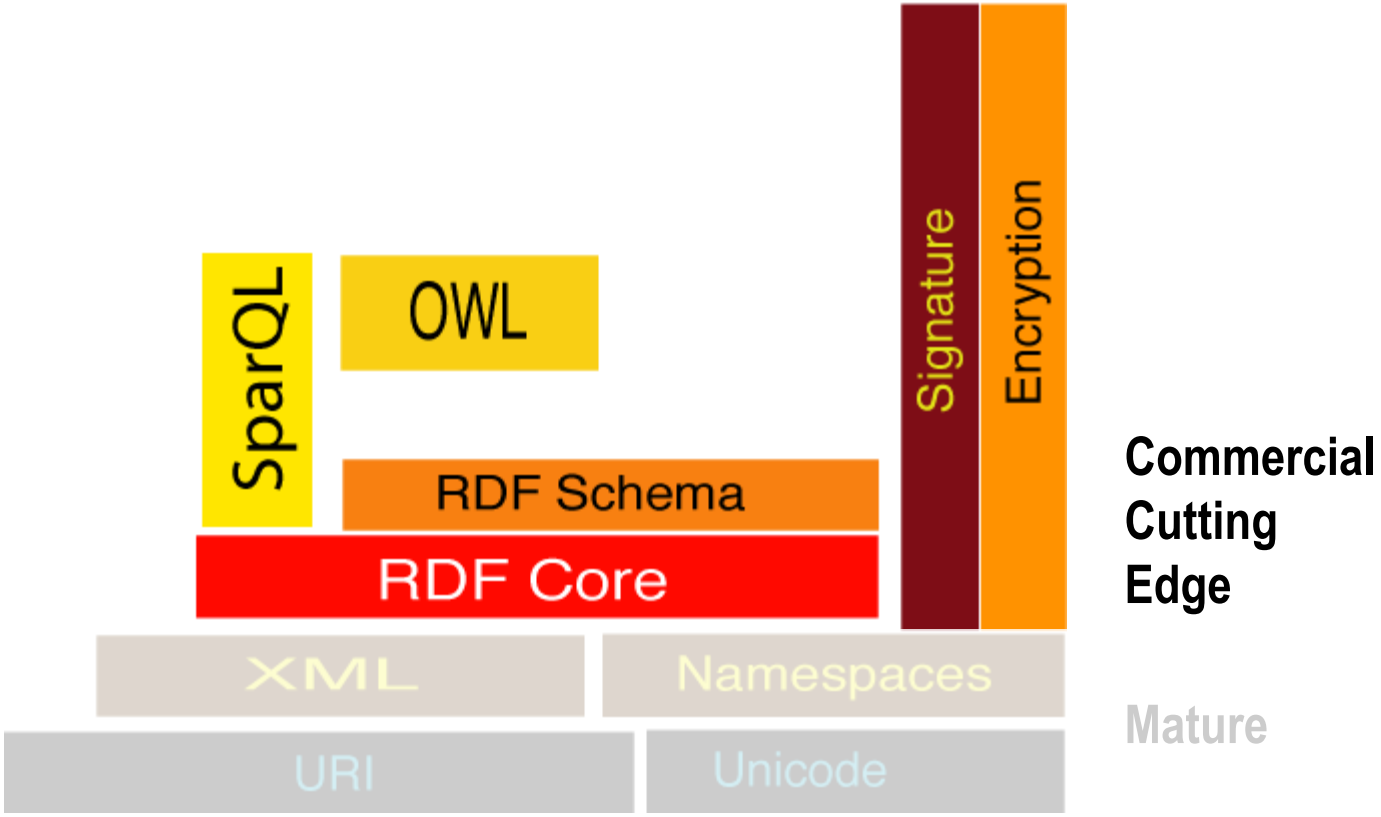
The Semantic Web in 2008

Mature

"The Famous Semantic Web Technology Stack"

VULCAN

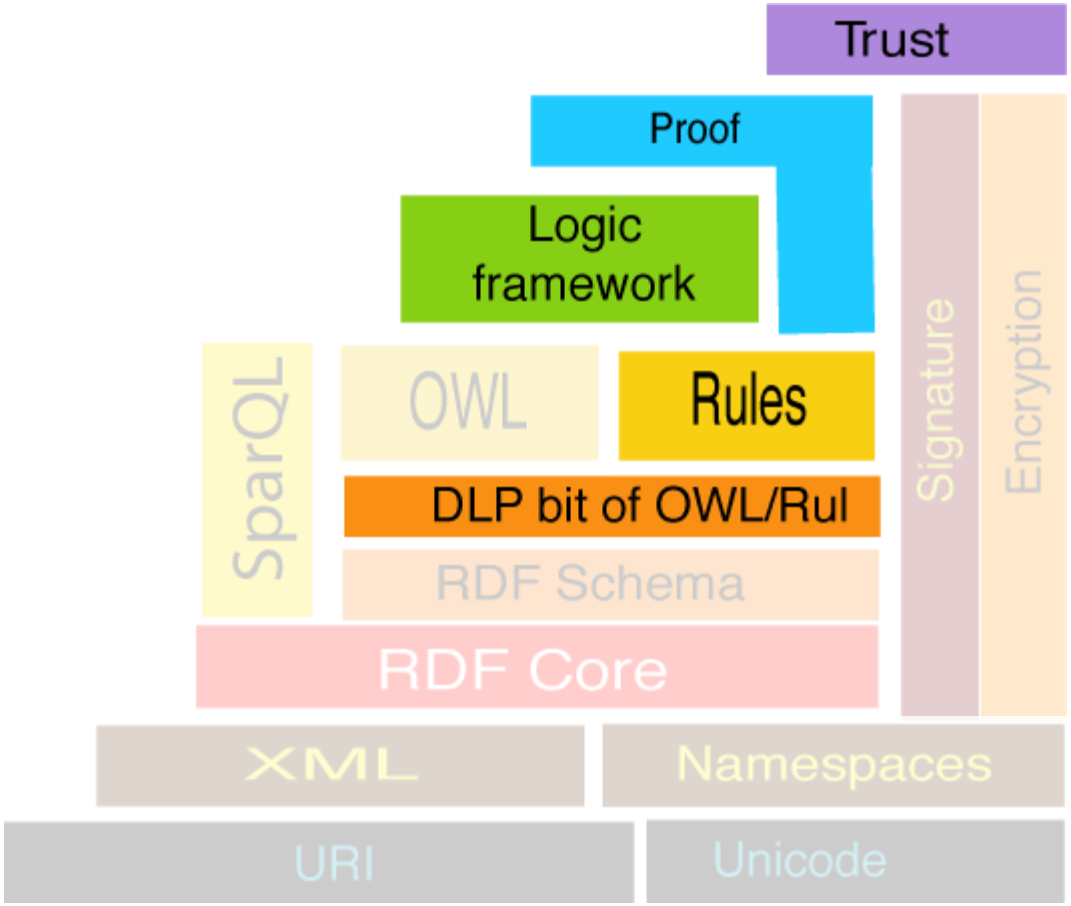
The Semantic Web in 2008



"The Famous Semantic Web Technology Stack"

VULCAN

The Semantic Web in 2008



Active Research and Standards Activity

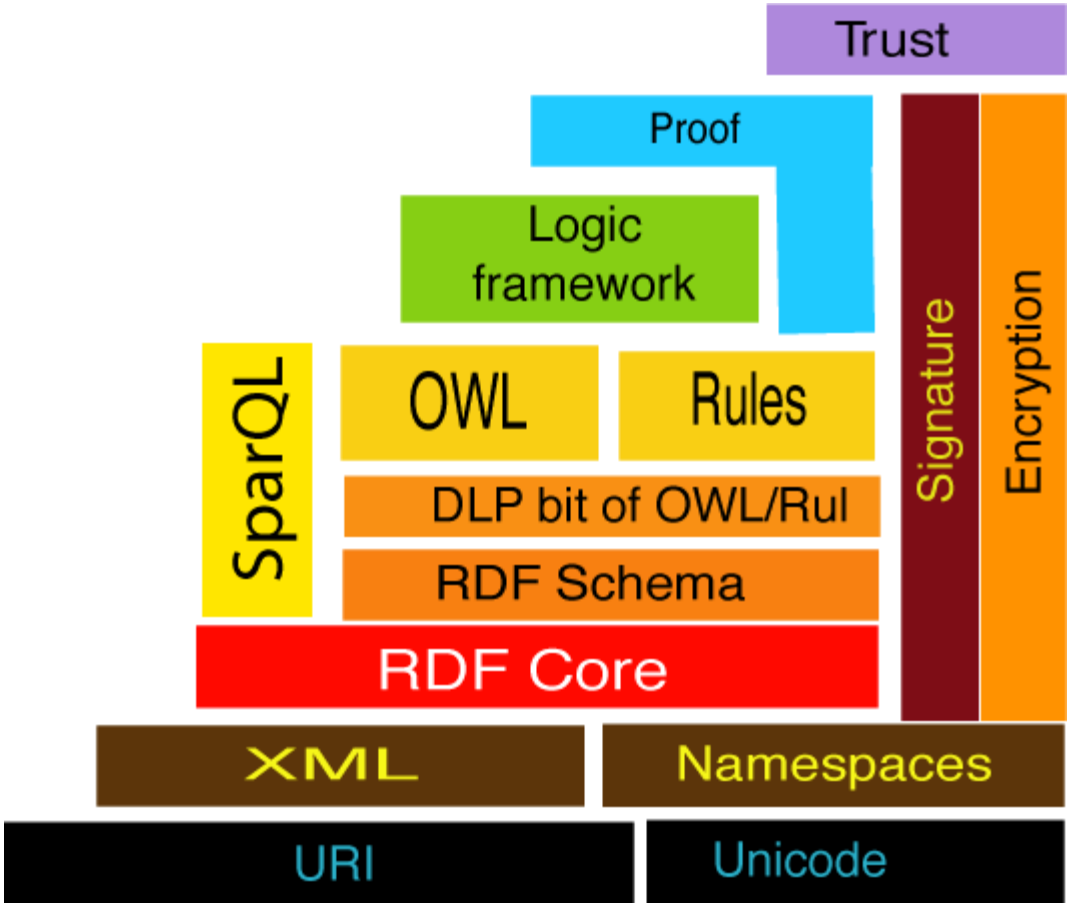
Commercial
Cutting
Edge

Mature

"The Famous Semantic Web Technology Stack"

VULCAN

The Semantic Web in 2008



Active Research and Standards Activity

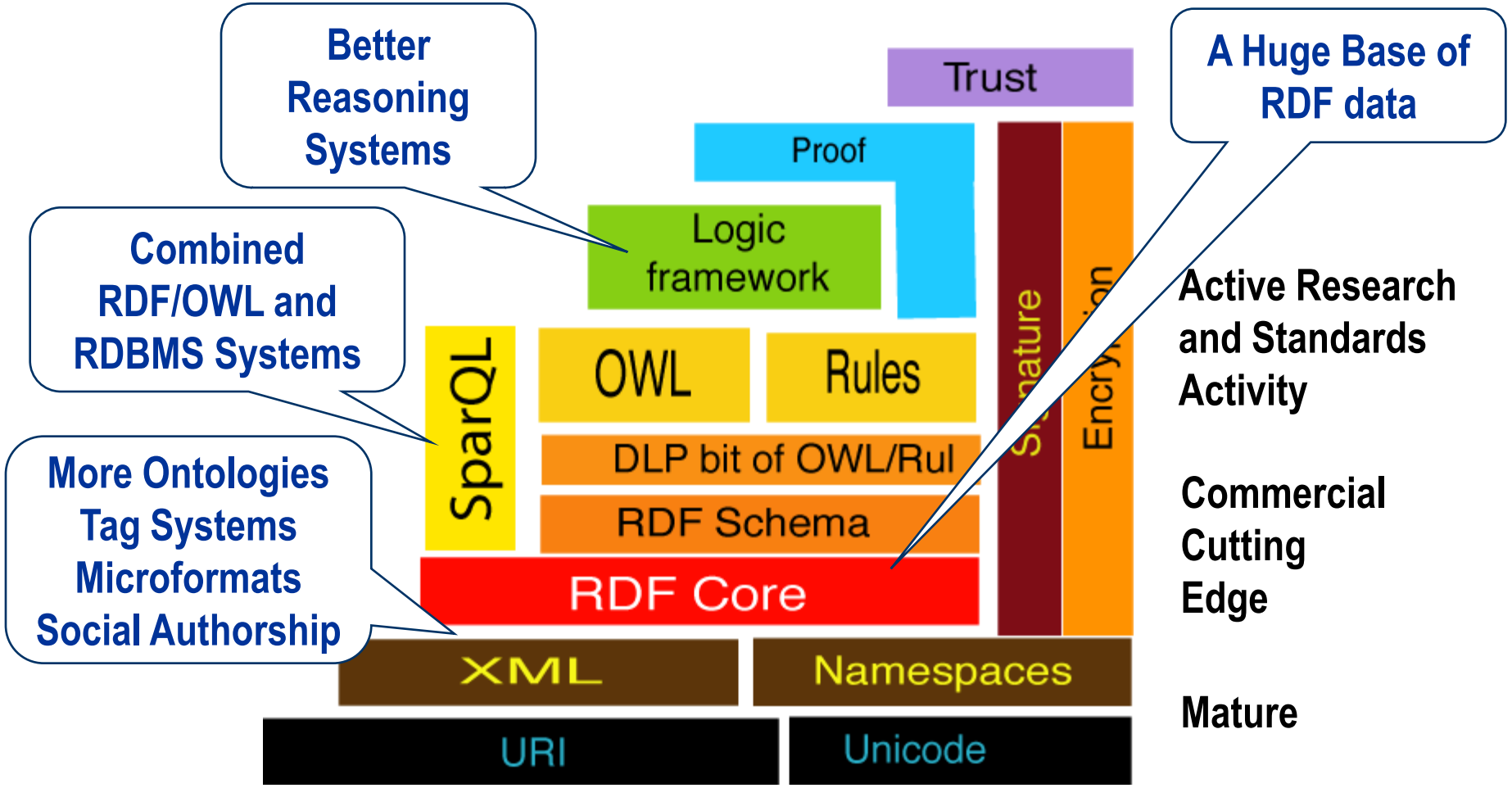
Commercial Cutting Edge

Mature

"The Famous Semantic Web Technology Stack"

VULCAN

Completing the Semantic Web Picture



Other Technologies Impact the Semantic Web

VULCAN

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 had a small semantic piece which is ending

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

- **An emerging commercialization ecosystem**
  - Startup: Radar, Metaweb, Evri, AdaptiveBlue, RealTravel...
  - Midsized: Monitor, Thetus, BBN, SRI, McDonald-Bradley, Franz...
  - Large: Yahoo!, Oracle, IBM, HP, Microsoft...
  - Semantic web exchange groups in Silicon Valley, Boston, Seattle...

- **Emphasis is mostly *Database* dimension of Semweb**
  - RDBMS scale and orientation, powerful analytics for Business Intelligence
  - Centralized workflows for ontology definition and management
  - Use cases surrounding data integration
  - Emerging microformats and structured blogging (e.g., Twine)
  - ... But mainly enterprise data description concerns

State of 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
    - €100M in 2007/8 and €70M in 2009/10 for Digital Libraries
    - Semantics is also present as a general systems technology
    - See http://cordis.europa.eu/fp7 for current investment amounts

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

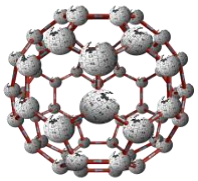
- **Emphasis on the *Social* and *Web* Dimensions of Semweb**
  - Web-scale, social networks, simple scalable imperfect inference
  - Ontology and data dynamics, imperfections, versioning
  - Semantically-boosted collaboration with limited knowledge engineer involvement
  - A base of socially-curated semantic data

Clear R&D leadership but lags in commercialization

Talk Outline: The Evolving Semantic Web

- **US Semantic Web R&D**
  - DARPA's DAML Program

- **Semantic Web Evolution to 2008**
  - Three Generations of Semantic Dreams
  - Markets and Companies



- **The Fourth Generation**
  - A Scalable Revolution

VULCAN

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 2008

- **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
  - 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 TBL

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
    - Documentum, Autonomy, Convera, FAST (bought by Microsoft for $1.2B)...
  - Email autoclassifiers and inbox managers
  - Web question answering
    - Hakia, Powerset, Answers.com, TextDigger, TrueKnowledge...
    - Cycorp
- **What are some issues with a VC bet in this space?**
  - Still waiting for a compelling match between technical capability and business need
    - Statistical methods are surprisingly good (e.g., Latent Semantic Indexing)
    - Verticals (esp. health care) 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 Example – 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
    - Questions like "What did Microsoft acquire in 2006?" or "What did Steve Jobs say about the iPod?"
    - No standard corpora to evaluate performance
    - Approaches Google in keyword search relevance performance
  - Plans to expand to multilingual and larger fragments of the web
  - Barney Pell (CEO) keynoted at ISWC 2007

- **Powerset's semantic knowledge is a superset of semantic web KR**
- **Bought by Microsoft in June for (rumored, approximate) USD$100M**
  - How was this valuation amount determined?
  - Fits with Microsoft strategy of search verticals
  - Reaction to Yahoo! Acquisition defeat

VULCAN

Second Generation Semantic Web Applications

Strategic Enterprise Information Technology

- **An only slightly newer problem type**
  - Exploitation of mainly structured enterprise data (RDBMS, Spreadsheets, files)
    - SOA integration, Enterprise Information Integration, Enterprise Application Integration
    - Backwards to Data Management to reduce cost of managing, migrating, integrating
    - Forwards to Business Process Management
  - Support for unified query, analytics, and application access
- **Markets Segments and Players**
  - Gardner estimates that EII software and services alone is $14B/year, with 40% growth over 5 years
  - Very complex market space includes EAI, Entity Analytics, MDM, BI, BPM, CPM...
  - Huge entrenched players (IBM, SAP, Oracle...) and consulting shops
  - Well-understood ROI and distribution models, well-resourced customers
- **What are some issues with a VC bet in this space?**
  - Fundamental problem is understanding the semantics from legacy systems, not in KR
  - Pure Semantic technology companies tend to be unsophisticated about customer business complexities
  - Tends to be an IT sale (not Line-of-Business sale), with attendant cost pressures

Semantic Submarkets in Strategic IT

	SDM	SBM	EII	BI	CPM	SOA	MDM	BPM
SDM	=	overlaps	is part of	is part of	is part of	is part of	is part of	is part of
SBM	overlaps	=		is part of	is part of			helps
EII	requires		=					
BI	requires	requires	requires	=				
CPM	requires	requires	requires	requires	=			
SOA	requires					=	is part of	is part of
MDM	requires		requires	helps	helps	requires	=	
BPM	requires					requires		=

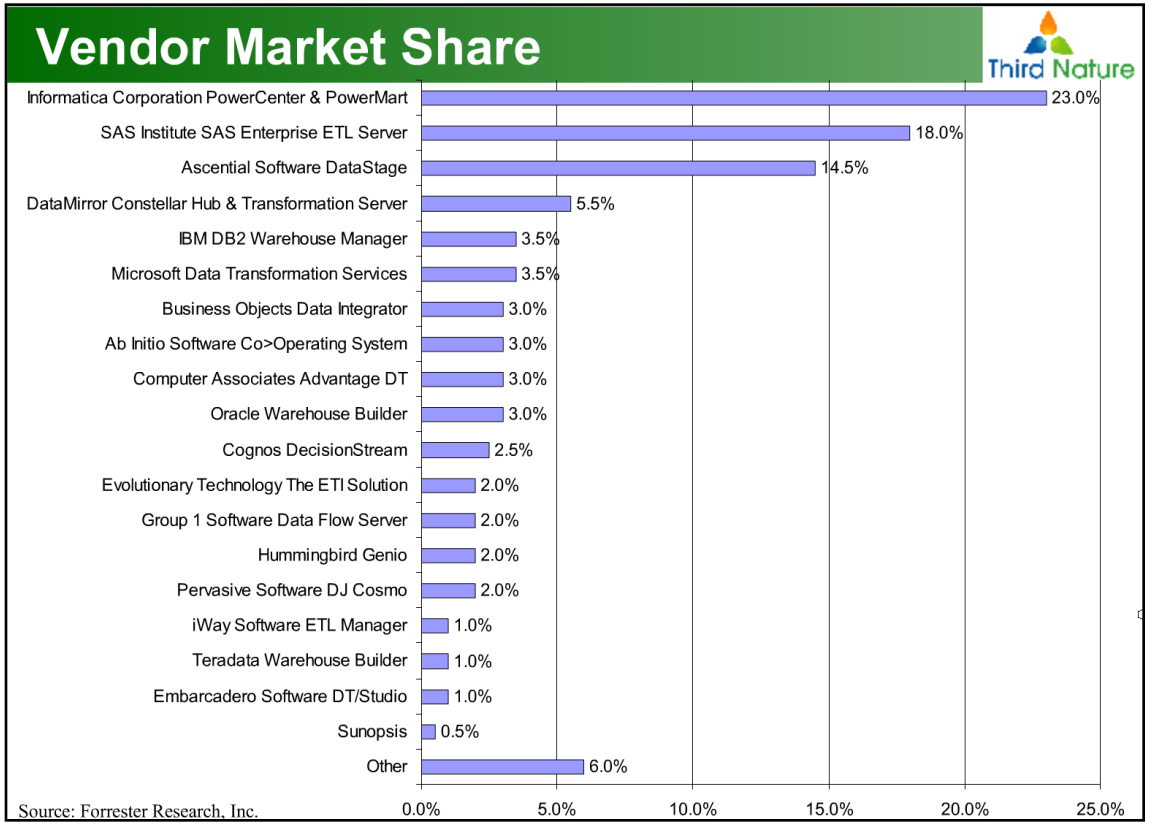
From Automata Inc.

- **Observations**
  - Wholesale reinvention is an impossible sale
  - Semantic Data Modeling is the critical core where we can get traction, but Semweb technology itself isn't a product
  - EII costs are mostly modeling
  - Most successful Strategic IT companies devote enormous resources to understanding the customer's data and outlook

Acronyms

SDM = Semantic Data Modeling
SBM = Semantic Business Modeling
EII = Enterprise Information Integration
BI = Business Intelligence
CPM = Corporate Performance Mgmt
SOA = Service Oriented Architecture
MDM = Master Data Mgmt
BPM = Business Process Mgmt

~$2B Market for EII Software



- Services are ~5x software; ~65% is “custom” work
- Good news is that the market is large
- Bad news is that it is not friendly to new technology startups

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, CEO of 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, Salzburgresearch, Ontoprise, Radar, ...)
  - Semantics in Social Networking (MySpace RDF service and microformats, Facebook, etc.)

- **What are some issues with a VC bet 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

Semantics without Semantic Web: Geocommons Mashups

- **How do Web 2.0 technologies currently share meaning?**
  - Tags, distinct social networks, group wikis
  - Explicit developer agreement on REST and SOAP and XQuery parameters
  - Mashup data fusion algorithms

- **Example: www.geocommons.com**
  - "The Hippest Places to Live in San Francisco"
  - Data sets used
    - San Francisco municipal neighborhood boundaries
    - Crime index by census tract
    - Home median age and density
    - Scraped ratings and locations of San Francisco bars and clubs
    - Occupation by census tract (techies and artists)
    - Commute mode (Female motorcycle ridership)

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: >2.5M English articles, >250M edits, >2.5M images, #8 Alexa traffic rank in August
  - 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://wiki.ontoprise.de for one powerful semantic wiki

Semantic Wiki Hypotheses:

(1) Significant interesting non-RDBMS Semantic Data can be collected cheaply

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

Third Generation Example: Metaweb and Freebase

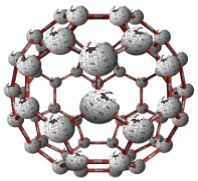
- Massive amounts of almanac-style RDF data (Creative Commons license) that is commonly available
- Social authoring tools and wiki-style consensus
- Data outsourcing model for long-tail startups

VULCAN

Talk Outline: The Evolving Semantic Web

- **US Semantic Web R&D**
  - DARPA's DAML Program

- **Semantic Web Evolution to 2008**
  - Three Generations of Semantic Dreams
  - Markets and Companies



- **The Fourth Generation**
  - A Scalable Revolution

VULCAN

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
    - Many of these involve natural semantically-oriented applications
  - 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

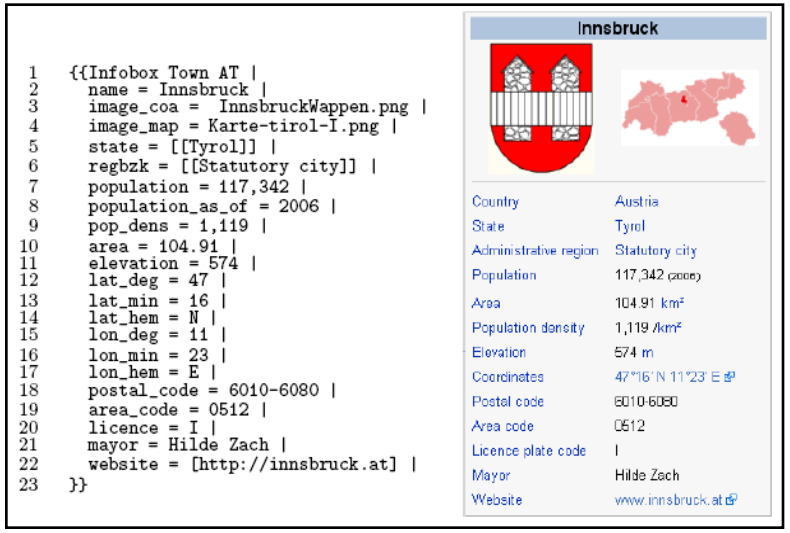
Material from Frank van Harmelen, Vrije Universiteit, Amsterdam

- **What are some issues with a VC bet 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?

VULCAN

Fourth Generation Example: DBpedia

- **Mine Wikipedia for assertions**
  - Mainly from Wikipedia Factboxes
    - ~23M triples
  - Category assertions
- **DBpedia 3.1 dataset (June 08 Wikipedia)**
  - ~2.5M things, ~218M triples
    - 108K persons, 392K places, 57K music albums, 36K films, 588K links to images, 3.1M links to relevant external web pages, 2.1M links into RDF datasets
  - Classifications via Wikipedia categories, YAGO, and WordNet synsets
  - One of the largest broad knowledge bases in the world
- **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

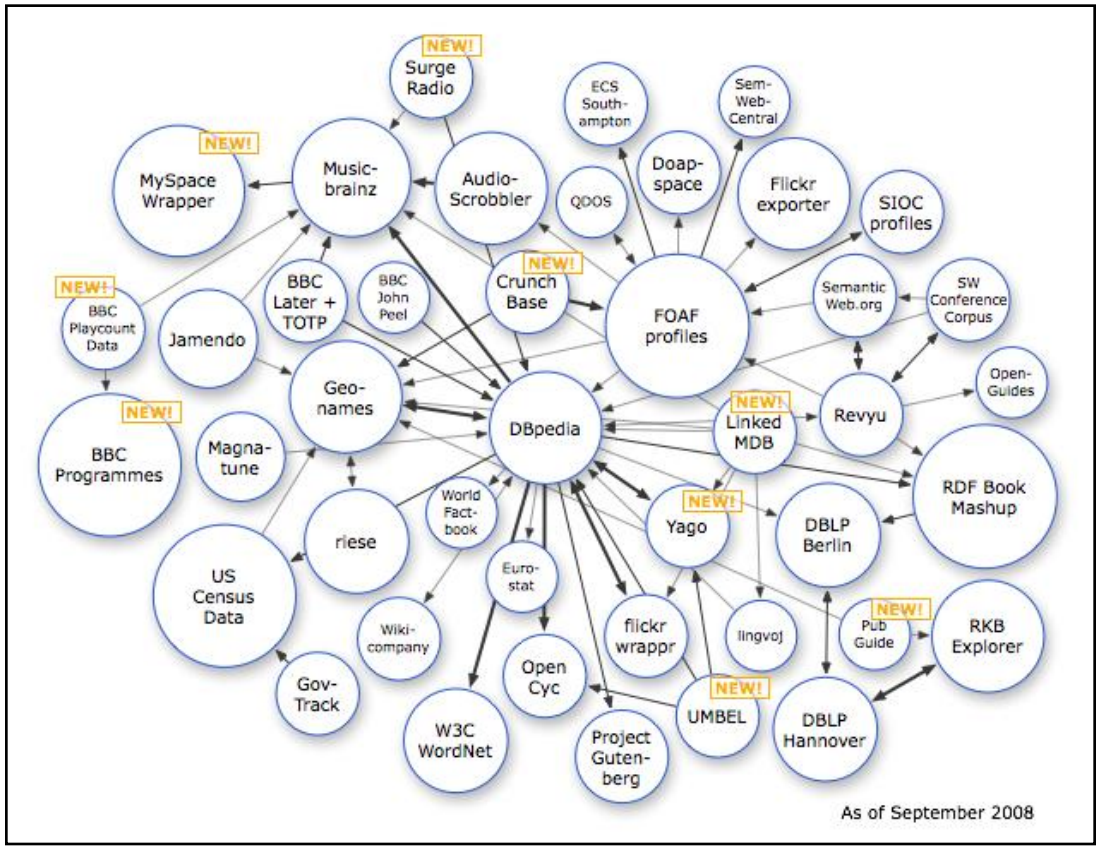
OPENLINK SOFTWARE

VULCAN

Fourth Generation Example: Linking Open Data

- **Goals**
  - Create a single, simple pub/access mechanism for web-scale 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
- **Total LOD dataset**
  - ~2B triples, and ~3B RDF links, and growing faster than I can track
  - Database linkage means that LOD will soon be impossible to count except via order of magnitude

LINKINGOPENDATA
W3C SWEO Community Project



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: >100K schemas
  - "Ontologies = Politics"

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

Material from Denny Vrandečić, AIFB

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**
  - Give up completeness
  - Combine reasoning/retrieval and search
  - Want to trade off answer quality and answer timeliness
  - Heavy emphasis on probability, decision theory, anytime algorithms
- **Reasoning pipeline**
  - 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
  - Plugin architecture, with sampling
  - Explicit cost models
- **The first time the semantic web has taken the *web* seriously**
  - 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 true *web-scale* KR**
  - answer timeliness
  - Heavy emphasis on probability, decision theory, anytime algorithms
- 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
  - Heavy emphasis on probability, decision theory, anytime algorithms
- Reasoning pipeline
  - Plugin architecture, with sampling
  - Kind of like SETI@Home

- **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™!***

VULCAN

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