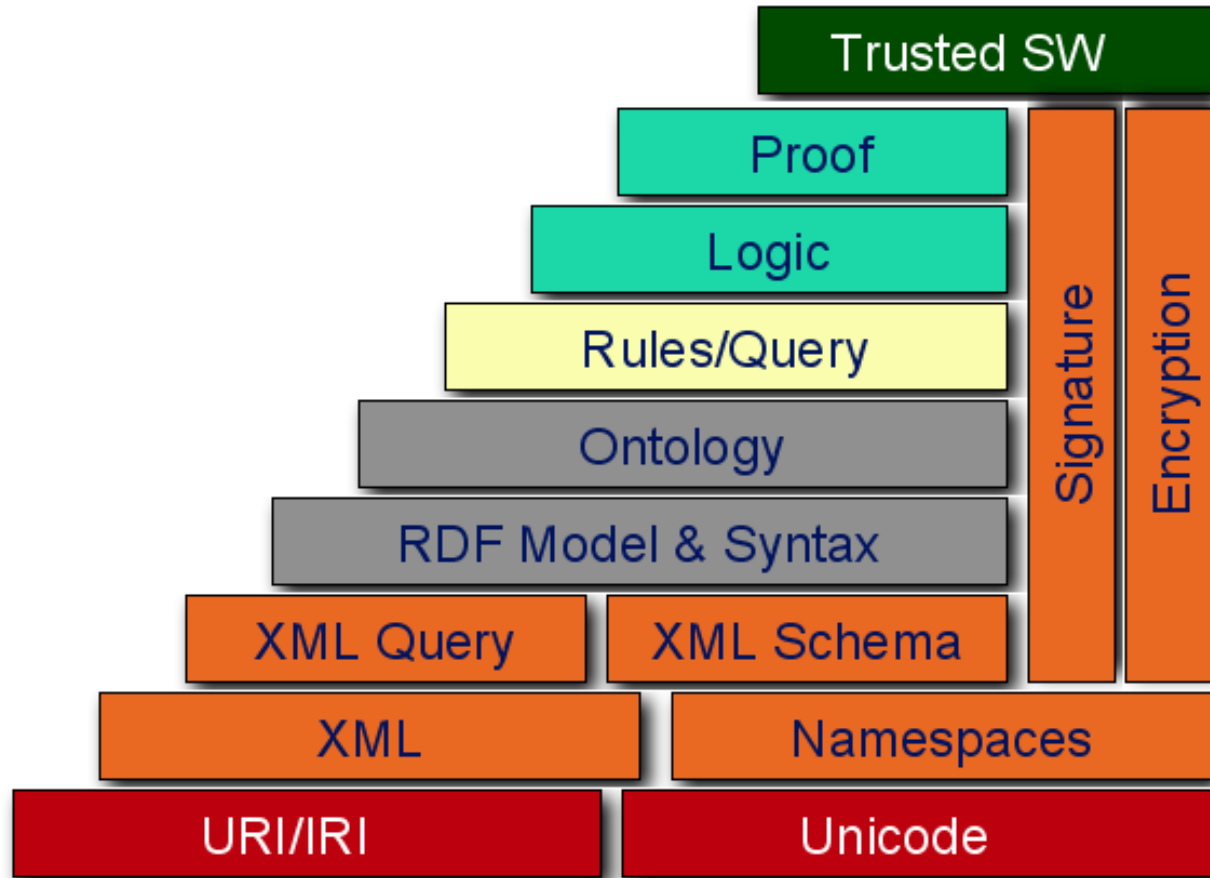


# A Pragmatic View of the Semantic Web and Ontologies

Mike Dean  
[mdean@bbn.com](mailto:mdean@bbn.com)

Opening Keynote  
[STIDS 2012](#)  
Fairfax, VA  
24 October 2012

# Are we done yet?



W3C, circa 2005

# Outline

---

- Accomplishments
- Current Work
- Surprises
- Predictions

# Accomplishments

---

- Standards
- Tools
- Linked Data
- Community
- Applications

# Standards

- Numerous Recommendations from W3C, ISO Common Logic, OMG ODM, OGC GeoSPARL
- Regularly revisited to incorporate user feedback and new technology

W3C Recommendation	Revisions
RDF	3
OWL	2
SPARQL	2

# Notable OWL 2 Extensions

---

- Property chains
  - Can express uncle as brother of parent
- Additional properties of properties
  - Reflexive, Irreflexive, Asymmetric
  - Useful to think about and record even if your reasoner doesn't yet support them
- Increased support for negative reasoning
  - Negative statements
  - Disjoint properties
- Profiles: EL, QL, RL

# Notable SPARQL 1.1 Extensions

---

- OWL 2 [Entailment Regimes](#)

# Tools

---

- Wide variety of high quality open source and commercial Semantic Web tools available
- Open source examples
  - [Protégé](#)
  - [Apache Jena](#)
  - [OWL API](#)
  - [D2RQ](#)
- Commercial examples
  - [TopBraid Suite](#)
  - [Pellet](#)
  - Oracle
  - IBM (DB2, Rational)
- [semwebcentral.org](#)
  - GForge instance with 166 open source software projects



# Triple Stores

---

- Huge improvement in the state of the art over the last 12 years
  - 10 billion statements on single server
  - Scalable distributed implementations
- Numerous providers
  - [AllegroGraph](#)
  - [OpenLink Virtuoso](#)
  - Oracle
  - IBM DB2
  - [OWLIM](#)
  - [Bigdata](#)
  - [Parliament](#)



# Linked Data Quality

---

- Overuse of owl:sameAs
- Better tool support
  - [LOD2 stack](#)
- Move toward authoritative sources
  - [data.ordnancesurvey.co.uk](http://data.ordnancesurvey.co.uk)
  - <http://cegis.usgs.gov/ontology.html>
  - [data.ign.fr](http://data.ign.fr)
  - [data.gov.uk](http://data.gov.uk)

# WikiData

---

- New Wikimedia Foundation project to populate infoboxes for all Wikipedia language versions from structured representations
- Essentially the inverse of DBpedia
- <http://www.wikidata.org>

# Linked Enterprise Data

---

- Apply [linked data principles](#) within organizational firewalls
- Optionally connected to public Linked Data cloud
- [Persistent URLs](#) map well to master data management
- Need to limit access to some enterprise data
  - I think HTTPS with client side certificate or password authentication is sufficient
  - Link to enterprise LDAP or other single sign-on solutions

# Community

---

- [W3C Semantic Web Activity](#)
- Conference Series
  - [ISWC](#)
  - [ESWC](#)
  - ASWC/CSWC/[JIST](#)
  - [FOIS](#)
  - [WWW Semantic Web Track](#)
  - [SemTech](#)
  - [STIDS](#)
- Virtual and local groups
  - [Ontolog Forum](#)
  - [Ontology Summit](#)
  - [Semantic Web Meetups](#)
- [Vocabulary camps](#)
- [International Association for Ontology and its Applications \(IAOA\)](#)

# Applications

---

- 1000s of Semantic Web applications in literature and in SemTech and other presentations
  - Not so many (yet) in the Apple or Android app stores
- Embedded use
  - IBM Watson
    - Chris Welty and others have given several talks on use of Semantic Web technologies and Linked Data in Watson
  - Apple Siri
    - Tom Gruber, David Martin, DARPA PAL heritage

# Recent Accomplishments

---

- GeoSPARQL
- RDF2RDB



# GeoSPARQL

- New Open Geospatial Consortium standard for representing and querying geospatial information
- Supports multiple
  - Geometries (points, lines, polygons)
  - Coordinate reference systems
  - Qualitative spatial relations (within, intersects, etc.)
- Preferred vocabulary for publishing new geospatial data
- <http://www.opengeospatial.org/standards/geosparql>
- [Parliament GeoSPARQL](#) is an open-source implementation

# RDB to RDF

---

- Much of the data on the (Semantic) Web resides in relational databases
- W3C has 2 new Recommendations for accessing such data
- [RDB to RDF Mapping Language \(R2RML\)](#)
  - Map a relational database to your own ontology
- [Direct Mapping of Relational Data to RDF](#)
  - RDF vocabulary automatically generated from database schema

# Ongoing Efforts

---

- SILK
- Provenance
- Ontology Design Patterns
- Ontology Repositories
- Earth Science
- Big Data
- Stream Processing
- Metrics and Quantification

# SILK

- Expressive rule language being developed by Vulcan, BBN, and others
- Semantic Web support
  - Import OWL 2
  - Import/export [RIF BLD](#) or [RIF SILK](#) dialects
- Rich enough to support policy and process modeling
  - Full support for negation
  - Prioritization
  - Justification
- Standards-maximizing development approach
  - Use a more expressive language only where needed
  - E.g. 50% OWL, 40% RIF, 10% SILK
- <http://silk.semwebcentral.org>

# Provenance

---

- Traceability of data from its source through various processing transformations is important
- W3C PROV addresses
  - Entities (e.g. documents), including Alternates
  - Activities (e.g. creation)
  - Agents (e.g. people, organizations, software)
  - Roles (e.g. editor)
  - Plans (e.g. workflows)
  - Derivation and Revision
  - Timestamps
- [http://www.w3.org/2011/prov/wiki/Main\\_Page](http://www.w3.org/2011/prov/wiki/Main_Page)
  - Start with the [PROV Primer](#)
  - Several documents are Last Call Working Drafts

# Ontology Design Patterns

---

- Analogous to software engineering design patterns
  - Promotes modularity and reuse of best practices
- Grew out of DOLCE modularization efforts
- Corresponding ISWC [Workshop on Ontology Patterns](#) series
- Several recent GeoVoCamps have focused on developing ODPs
- <http://ontologydesignpatterns.org>

# Ontology Repositories

---

- [NCBO BioPortal](#) has been highly successful with a large user base
- Open Ontology Repository initiative
  - Focus of [Ontology Summit 2008](#)
  - Great collaboration with BioPortal, Toronto, Bremen, and other groups
  - Still primarily a volunteer effort
  - Limited progress
- Collection of (mostly BioPortal-based) repositories
  - e.g. [socop.oor.net](#)

# Earth Science

---

- Community seems poised to become a major adopter of semantic technologies, a la Health Care and Life Sciences
- [NSF EarthCube](#) program
  - [Semantics and Ontologies group](#) has 113 members
- Builds off work of Peter Fox and Deb McGuinness, the late Rob Raskin, [SOCoP](#), et al.
- [EarthScienceOntolog](#) mini-series



# Big Data

---

- Potentially big opportunity for Semantic Web and ontologies
- Need to expose data set models/vocabularies/ontologies
- Support both data and metadata (registries)
- [RDF Data Cube Vocabulary](#)
  - Vocabulary for publishing multi-dimensional data, such as statistics, as Linked Data
  - Supports units of measure and slices
  - Developed for data.gov.uk
  - Extend to support “standoff annotation” of large data sets

# Stream Processing

---

- Stream processing of Semantic Web content offers scalability, efficiency, and latency advantages
- Knowledge Streams concept presented at SemTech DC 2011
- EU [Large Knowledge Collider \(LarKC\)](#) project employs stream based reasoning
- Several groups now working on Semantic Complex Event Processing

# Metrics and Quantification

---

- SSWS 2008 invited talk on Towards a Science of Knowledge Base Performance Analysis
  - Included an analysis of the Billion Triples Challenge Corpus
  - Related to Frank van Harmelen's ISWC 2011 keynote on searching for universal patterns
- Seeking a metric for measuring the richness of an ontology
  - Beyond more qualitative DL expressivity, e.g. SROIQ(D), or graph measures

# Semantic Information Theory

---

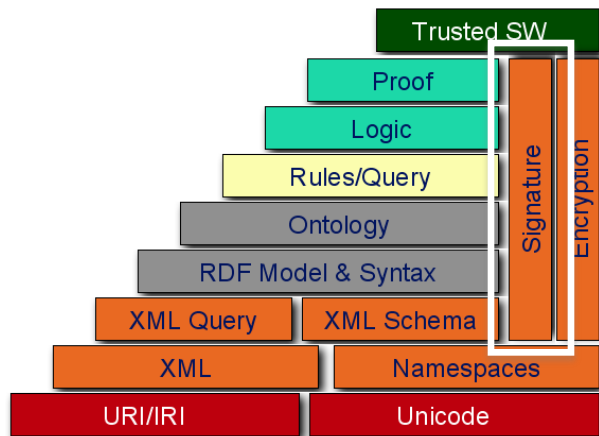
- Task with Jim Hendler and others in the Army Research Lab [Network Science Collaborative Technology Alliance](#)
- Addressing semantics and utility explicitly ignored by Shannon
- Theoretical underpinnings presented at [NSW 2011](#) and [IQ2S 2012](#)
- Paper on Utility in the Semantic Web to be presented in the [Quantitative Formalization in the Semantic Web](#) workshop at ISWC 2012

# Surprises

---

- Digitally Signatures
- Probabilistic Reasoning

# Digital Signatures



- I've never seen a digitally signed RDF document in the wild
- My colleague Doug Reid showed years ago that this was doable
- Need support within tools layer
- Perhaps reflects relatively slow uptake of PKI in general

# Probabilistic Reasoning

---

- When many people think about the Semantic Web, they focus on inference
- I think most interesting inferences are probabilistic rather than logical
- Despite on-going research by various groups, there's still little consensus on how to combine logical and probabilistic reasoning

# Predictions

---

- Continued Growth
- R.V. Guha's Semantic Trajectory
- Structured Reporting



# Continued growth

---

- Use of Semantic Web technologies will continue to increase at current or accelerating rates
- No obvious replacement on the horizon

# R.V. Guha's Semantic Trajectory

Timeframe	Knowledge Representation
1987-1994	Cyc
1994-1997	MCF (an RDF precursor)
1997-1999	RDF
2000-2002	TAP
2005-present	schema.org

- [Guha](#) is a very smart guy who's been working in our field a long time. There's a good chance he's ahead of us.
- schema.org is compatible with the Semantic Web and has been widely adopted by the mainstream web
  - Panel at SemTech 2012 in San Francisco
- After years of “dumbing down” the KR, perhaps we can get enough traction with schema.org to begin moving the mainstream web toward more expressive representations

# Structured Reporting

---

- Years ago I semi-joked that we'd see the Columbia School of Journalism and Ontology
- We spend a significant portion of 20+ years of kindergarten through doctoral or other professional education teaching (with varying success) written natural language communication, but are generally unwilling to invest a small fraction of that time teaching structured knowledge representation
- Despite incremental progress, reliable extraction of facts from text remains a research problem
- I believe that with proper tools and training, professionals can enter facts directly (often at lower or comparable cost)
- Possible technology approaches
  - Controlled natural language
  - Improved user interfaces
  - Event-specific templates or apps
- Alternatively, this could become a specialized clerical skill like stenography or court reporting

# Upcoming Events

---

- 11<sup>th</sup> International Semantic Web Conference
  - Boston, November 11-15
  - <http://iswc2012.semanticweb.org>
- SOCoP Workshop
  - USGS Reston, November 29-30
  - Follow-on to GeoVoCamps in [DC](#), [Santa Barbara](#), and [Dayton](#)

# Questions?

---