Semantic Integration with Apache Jena and Apache Stanbol

All Things Open

Raleigh, NC

Oct. 22, 2014

Overview

- Theory (~10 mins)
- Application Examples (~10 mins)
- Technical Details (~25 mins)

What do we mean by “Semantic Integration”?

- Integration, generally
- Letting things “talk to each other” so they can act as a cohesive whole
- Uses the Semantic Web technology stack
- Data integration using RDF, well known vocabularies, as well as in-house vocabularies and ontologies.
- Relationship to EAI, MDM, etc?

Uses Semantic Web technology to do what, exactly?

- Work with knowledge, not labels
- Express metadata about “things”
- And the relationships between those “things” and their characteristics
- Reason about those “things” in order to:
  - Find contextually relevant information
  - Search with greater precision
  - Generate new knowledge
  - ???

Knowledge?

- What's the difference between “Data”, “Information”, “Knowledge”, etc?
- Different ways of talking about this.
- DIKW Pyramid is a popular model
- http://en.wikipedia.org/wiki/DIKW_Pyramid


Wisdom
Knowledge
Information
Data

Knowledge?

- For our purposes today...
  - Unambigous Identifiers
  - Ontology
    - Type / Class information
    - Relationships

Working With Knowledge instead of Labels

- Backing up – what do we mean by “Semantic” anyway?
- Is “Java”:
  - An island in the South Pacific
  - A slang word for coffee
  - A programming language invented by Sun Microsystems
- Using URIs as labels
  - In order to talk about “the semantics of Java” we have to know unambiguously **which** “java” we are talking about.

Ontology

- The attributes / properties of a Thing
- Set membership of a Thing
  - rdfs:Class
- Relationships between Things
  - dc:relation
  - dc:subject
  - rdfs:subClassOf
  - skos:narrower, skos:broader

Data Table Slide

id	color	size	manufacturer
2345	Blue	Large	Acme
2378	Red	Small	Cullet
3421	Green	Medium	Acme

Data as Triples

subject	predicate	object
uid:2345	rdf:type	owl:Thing
uid:2378	rdf:type	owl:Thing
uid:3421	rdf:type	owl:Thing
uid:2345	pref:color	“Blue”
uid:2378	pref:color	“Red”
uid:3421	pref:color	“Green”
uid:2345	pref:size	“Large”
uid:2378	pref:size	“Small”
uid:3421	pref:size	“Medium”
uid:2345	pref:manufacturer	uid:9998
uid:2378	pref:manufacturer	uid:9997
uid:3421	pref:manufacturer	uid:9998
uid:9998	rdfs:label	“Acme”

Types & Relationships

- RDF/S
  - superclass / subclass relationships for Classes
  - superclass / subclass relationships for Properties
  - domain / range relationship between Properties and Classes
- OWL
  - class equivalence
  - entity equivalence
  - class disjointness
- SKOS
  - narrower / broader relationship between Concepts
  - ordered collections

But

- But... we're not here for a course on Epistemology or Metaphysics...

Synonyms

- Smart Data
- Semantic Data
- Knowledge

Semantic Integration Layer

Enterprise Applications
(ERP, SFA,
CRM, etc.)

Document Repositories
DMS, Wikis, Blogs,
Forums, Etc.

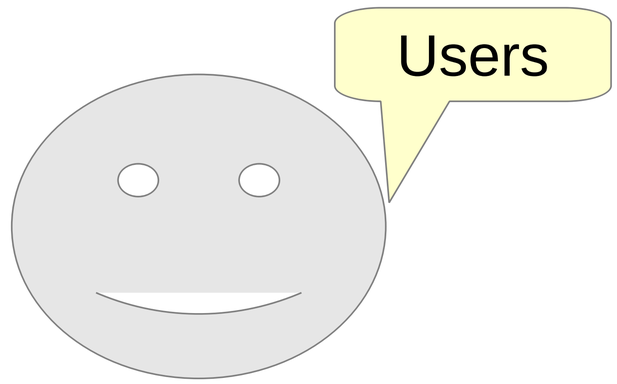
“Big Data”
Data Warehouses,
Data Lakes, etc.

Stanbol

Jena

Internet of Things,
M2M, Sensor Data
etc.

“Open Data”
SEC filings
EPA data
building permits,
etc.



But wait, there's more...

- From relational database to Semantic Web -> R2RML
  - D2RQ
  - http://d2rq.org
- ANY23 – Anything to Triples
  - http://any23.apache.org
- OpenRefine, Tika, JSoup, Boilerpipe, ...
- Potentially, anything that might be part of a normal ETL workflow

So, what is the Semantic Web?

Sir Tim Berners-Lee's vision of the Web as a universal medium for data, information, and knowledge exchange.

An evolving extension of the World Wide Web in which the semantics of information and services on the web is defined, making it possible for the web to understand and satisfy the requests of people and machines to use the web content.

...prospective future possibilities that are yet to be implemented or realized.

A set of design principles, collaborative working groups, and a variety of enabling technologies.

What is the Semantic Web? (continued)

"*... supposed to make data located anywhere on the Web accessible and understandable, both to people and to machines.*"

(Explorers Guide to the Semantic Web, p 3)

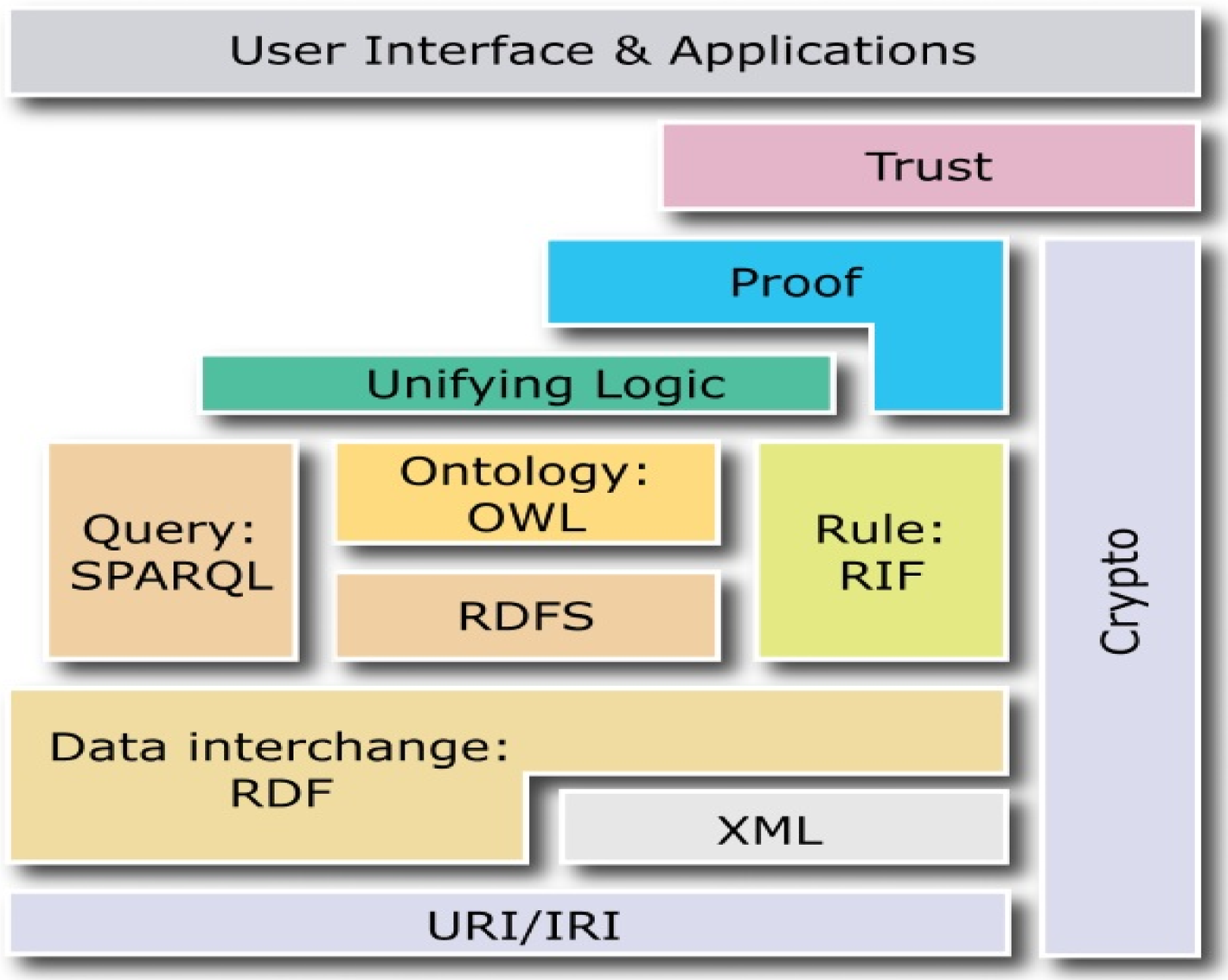
"*... more a vision than a technology.*"

(Explorers Guide to the Semantic Web, p 3)

"*...a fluid, evolving, informally defined concept rather than an integrated, working system.*"

(Explorers Guide to the Semantic Web, p 3)

The “Semantic Web Layer Cake”



RDF – Resource Description Framework

Models statements as “triples”

Subject, Predicate, Object

- Resources unambiguously named using URIs
- Everything is a triple... ex: “the shoe is red” would be the triple with subject = “shoe”, predicate (or property) = “color”, and object (or value = “red”
- Serialization formats include XML (known as RDF/XML) and developer friendly serialization formats including N3, Turtle, and JSON-LD

Reasoning over data

- OWL / SKOS / etc.
- Ability to access “Inferred” triples

Common Vocabularies

- FOAF
- SKOS
- DOAP
- Dublin Core
- Etc.

Querying with SPARQL

- Basic queries
- Using inferred triples
- Federated Queries
- DBPedia example

Semantic Integration in the Enterprise

- Knowledge Management
- Collaboration
- BPM
- Business Intelligence
- Predictive Analytics

Apache Jena

- RDF API
- Triplestore (TDB)
- Sparql Execution Engine (ARQ)
- OWL Reasoner
- SPARQL endpoint (Fuseki)
- Inference API
  - Use built in reasoners
  - Or define your own inference rules
- http://jena.apache.org

Apache Stanbol

- A “RESTful Semantic Processing Engine”
- Use cases
  - Content Enhancement
  - see:
    - http://stanbol.apache.org/docs/trunk/scenarios.html
  - ContentHub, EntityHub, etc.
  - Quoddy scenario demo
- http://stanbol.apache.org

Not AI, but...

- Newer reasoners can utilize new techniques, including Bayesian inference, any sort of machine learning models, cognitive models, new NLP techniques, etc.
- Same for Stanbol extraction – you can write your own extractors and new extractors will be coming down the pipe.