Semantic Web and the Post Relational World

DAMA Phoenix

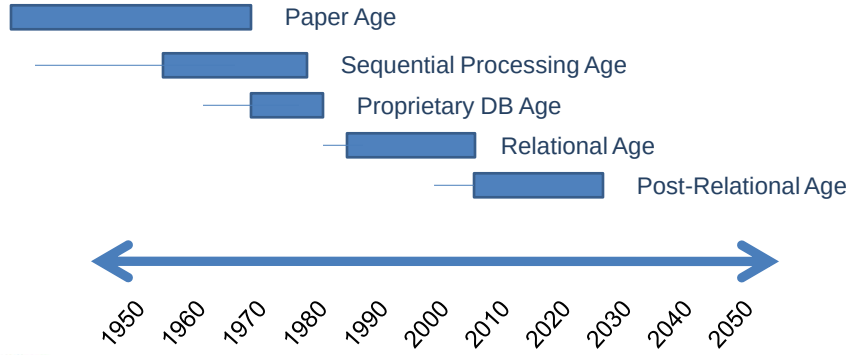
June 17, 2010

Dave McComb

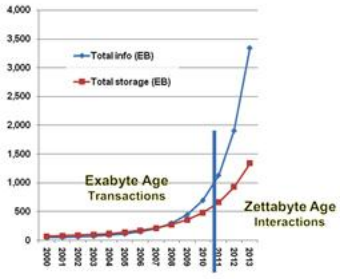
Semantic Arts

The "Post-Relational" World

Time line



These days, we are drowning in data



www.c-data.nl

We have created more data since 2005
than we have in the previous 40,000 years.

Limits of Relational

Storage — Google doesn't store it's data in relational databases. There isn't a database that could handle it

Performance — There are many DBMS's now that outperform relational by 50x (Hadoop, Mapreduce, Columnar DBs etc)

Complexity — Relational databases rely on metadata to create and communicate distinctions.

Sir Tim Berners-Lee

Web 3.0/ Semantic Web

- What is it, and how does it turn data into information and knowledge?

Three key distinctions in Web 3.0 (aka the Semantic Web)

- A uniform way to refer to specific instances.
- One way to declare all facts.
- A way to describe classes/categories/types/sets in a way that allows computers to categorize some data for us.

Instances

- Anything individually identifiable
  - People
  - Organizations
  - Cars
  - Documents
  - Contracts
  - Transactions
  - Etc.

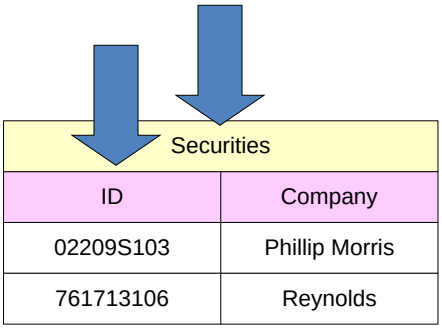
Historic Treatment

- Create a table
- Give it an “id” or “key”

Securities	
ID	Company
02209S103	Phillip Morris
761713106	Reynolds

Historic Treatment

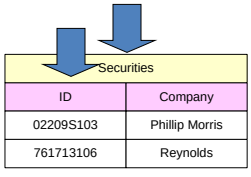
- To get data out, you need to know the table and the column (the metadata).



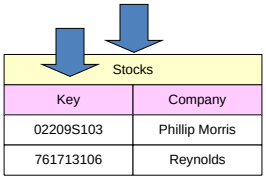
Securities	
ID	Company
02209S103	Phillip Morris
761713106	Reynolds

Historic Treatment

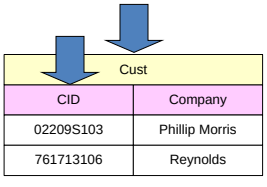
- To get data out of dozens of systems, you'd need to know dozens of bits of metadata.



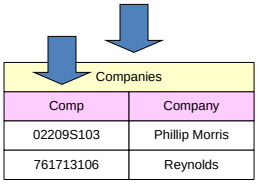
Securities	
ID	Company
02209S103	Phillip Morris
761713106	Reynolds



Stocks	
Key	Company
02209S103	Phillip Morris
761713106	Reynolds



Cust	
CID	Company
02209S103	Phillip Morris
761713106	Reynolds



Companies	
Comp	Company
02209S103	Phillip Morris
761713106	Reynolds

This doesn't scale

Web 3.0 approach

- Skip the metadata (for identification).
- Everything is a resource.
- Everything gets a "URI" (think URL).

cusip:02209S103

http://www.cusip.com

I know what you're thinking…

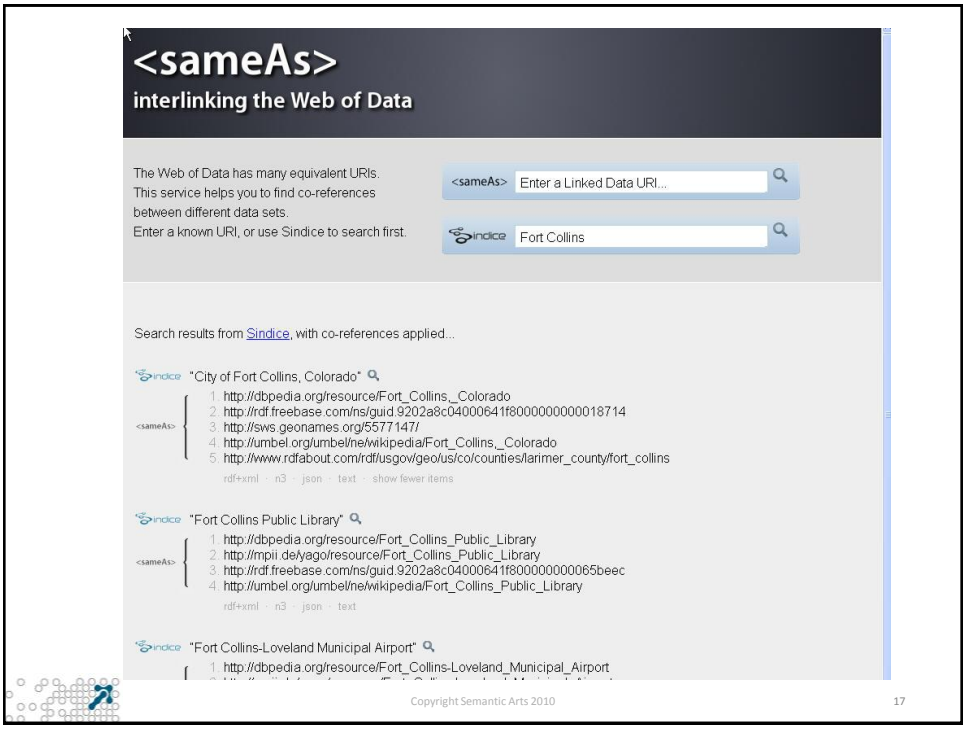
Does this mean everyone has to use the same id?

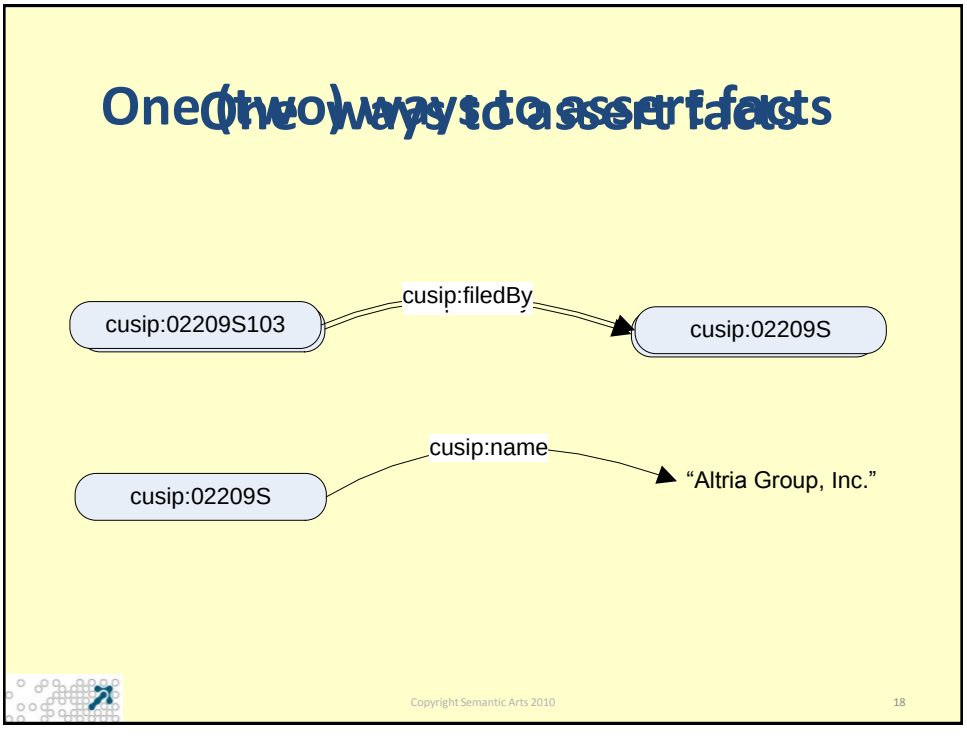
Won't this be like every "universal id" system we've ever seen?

Nope.

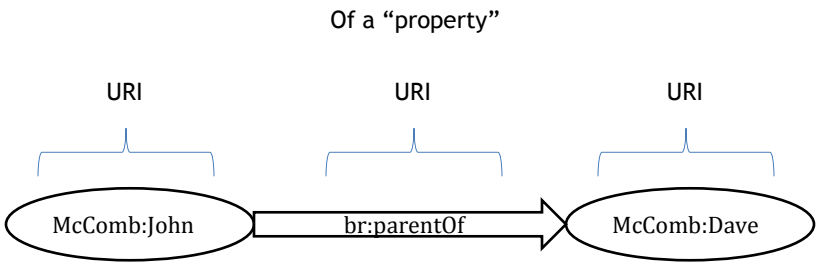
This doesn't require that everyone use the same id, only that if you use the id you refer to the same thing.

What if two systems assign different identifiers to the same thing?


<sameAs>
interlinking the Web of Data
The Web of Data has many equivalent URIs.
This service helps you to find co-references
between different data sets.
Enter a known URI, or use Sindice to search first.
<sameAs> Enter a Linked Data URI...
Sindice Fort Collins
Search results from Sindice, with co-references applied...
Sindice "City of Fort Collins, Colorado"
<sameAs>
1. http://dbpedia.org/resource/Fort_Collins,_Colorado
2. http://rdf.freebase.com/ns/guid.9202a8c04000641f8000000000018714
3. http://sws.geonames.org/5577147/
4. http://umbel.org/umbel/ne/wikipedia/Fort_Collins,_Colorado
5. http://www.rdfabout.com/rdf/usgov/geo/us/co/counties/larimer_county/fort_collins
rdf+xml · n3 · json · text · show fewer items
Sindice "Fort Collins Public Library"
<sameAs>
1. http://dbpedia.org/resource/Fort_Collins_Public_Library
2. http://mpii.de/yago/resource/Fort_Collins_Public_Library
3. http://rdf.freebase.com/ns/guid.9202a8c04000641f8000000000065beec
4. http://umbel.org/umbel/ne/wikipedia/Fort_Collins_Public_Library
rdf+xml · n3 · json · text
Sindice "Fort Collins-Loveland Municipal Airport"
1. http://dbpedia.org/resource/Fort_Collins-Loveland_Municipal_Airport
Copyright Semantic Arts 2010
17


One (two) ways to assert facts
cusip:02209S103
cusip:filedBy
cusip:02209S
cusip:name
cusip:02209S
“Altria Group, Inc.”
Copyright Semantic Arts 2010
18

Anatomy of a triple

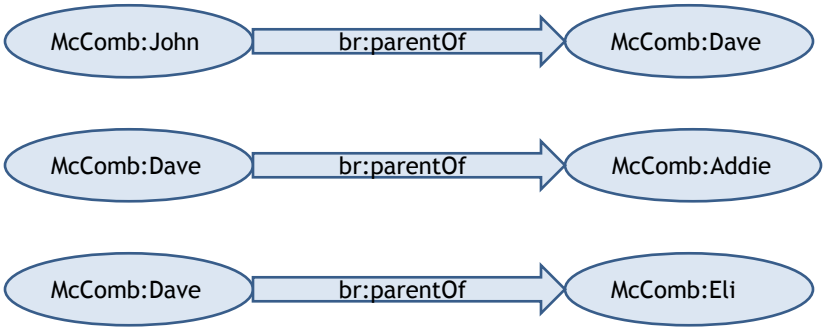


The Logical Model of the Semantic Web

Subject	Predicate	Object
URI	URI	URI
URI	URI	Literal

Triples

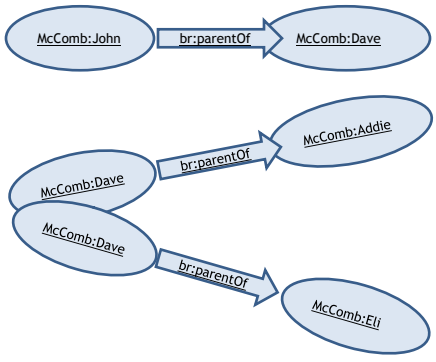
Subject	Predicate	Object
McComb:John	br:parentOf	McComb:Dave
McComb:Dave	br:parentOf	McComb:Addie
McComb:Dave	br:parentOf	McComb:Eli



Here's where the magic is....

Triples to Graphs

Linked Data



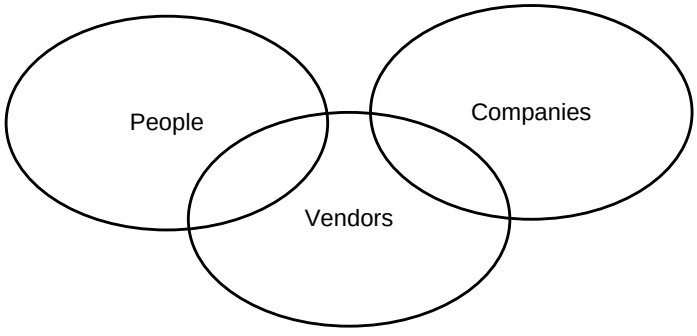
The "join" is done at the instance level

- No metadata was harmed in the making of this join.
- The metadata wasn't even interrogated.
- In fact, it just isn't necessary.

Schema

- In traditional systems, “schema” defines physical structure as well as hinting at meaning, and must be defined before data can be stored.
- In “Semantic-land” the schema is “logical” (not physical) and “late” (can be bound after the instances have been created).

Classes, Categories, Sets, Types



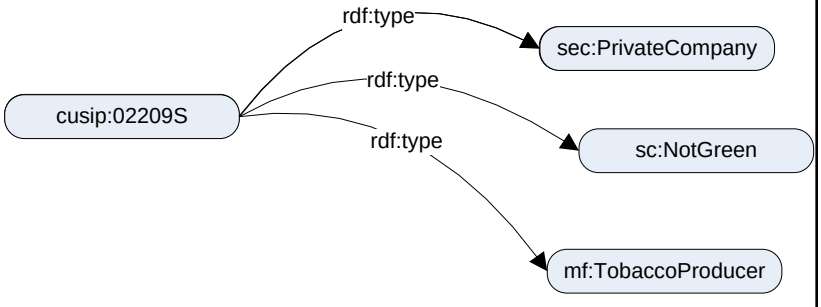
Current Approaches

- Humans assign things to categories.
- Things get one primary category and that category's parents.
- Once assigned, items stay in their categories.

Web 3.0 approach

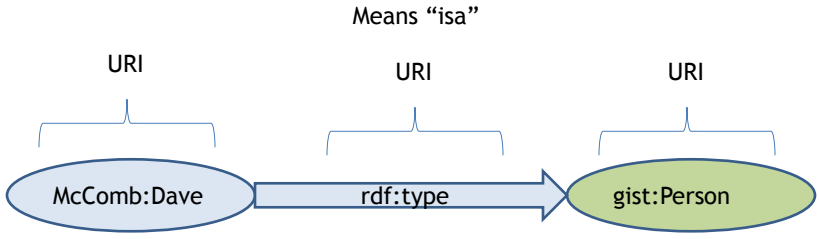
"Type" or "Class" is not structural. It's just another assertion.

Any instance may be many types simultaneously.



Using Existing Classes

Subject	Predicate	Object
McComb:Dave	rdf:type	gist:Person



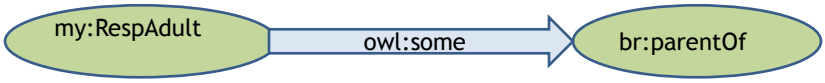
Making Up Classes (schema later and local)

Subject	Predicate	Object
my:RespAdult	owl:some	br:parentOf

Means: "you are a member of the class of Responsible Adults, if you are the parent of at least one [Person]"

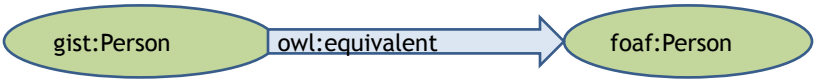
Inference

Subject	Predicate	Object
McComb:John	br:parentOf	McComb:Dave
McComb:Dave	br:parentOf	McComb:Addie
McComb:Dave	br:parentOf	McComb:Eli
McComb:John	rdf:type	my:RespAdult
McComb:Dave	rdf:type	my:RespAdult



Integration

Subject	Predicate	Object
gist:Person	owl:equivalent	foaf:Person

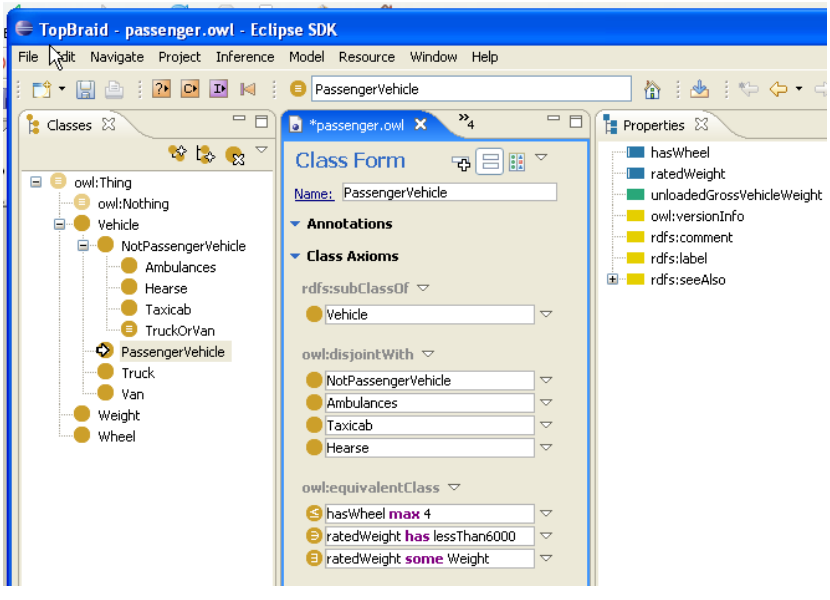


Means: “any one who is a ‘gist:Person’ is also a ‘foaf:Person’ and vice versa.

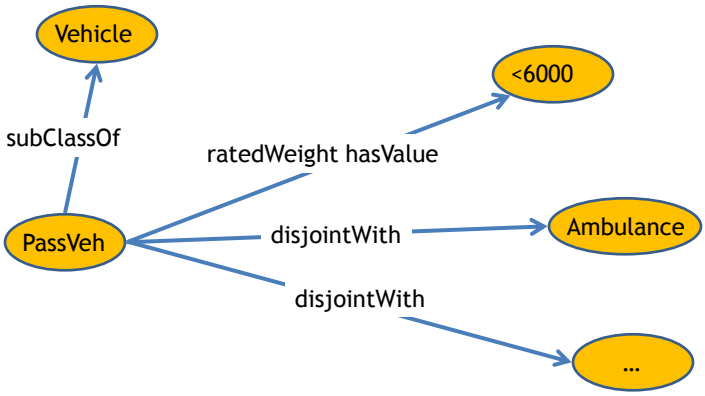
IRS definition of a passenger automobile

- A passenger automobile is a 4-wheeled vehicle manufactured primarily for use on public roads that is rated at 6,000 pounds unloaded gross vehicle weight or less. Certain vehicles, such as ambulances, hearses, and taxicabs, are not considered passenger automobiles and are not subject to the line 36 limits…

  Form 4797

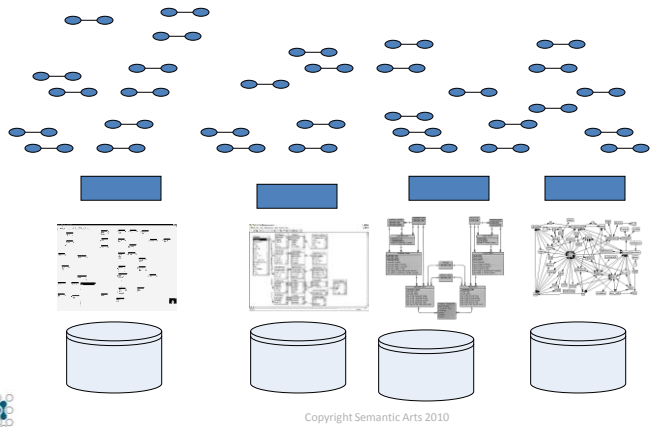


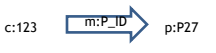
Even the definition of classes is done in triples



So where are these "triples" coming from ?

Triples from Databases

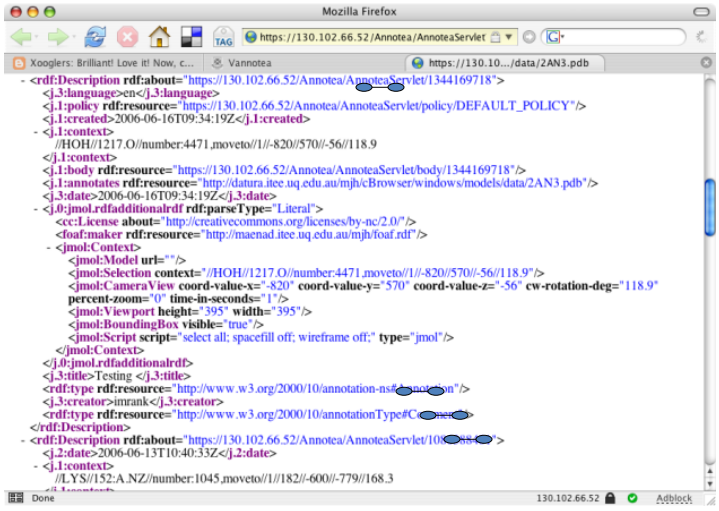




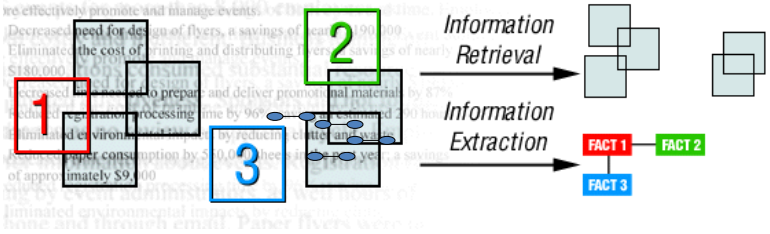
C_ID	O_ID	P_ID	Qty	Price
123	AA001	P27	4	1.00
456	AA002	P27	6	1.00

Prod	Desc	OnHand
P27	Bolt	55
P28	Nut	66

Triples from XML/HTML



Triples from Text



Aerotext from Lockheed Martin

Triples and Triple Stores

- Once we've harvested a large number of "triples" we need a place to store and efficiently access them.
- This is the role of a "triple store," essentially a database for these assertions.

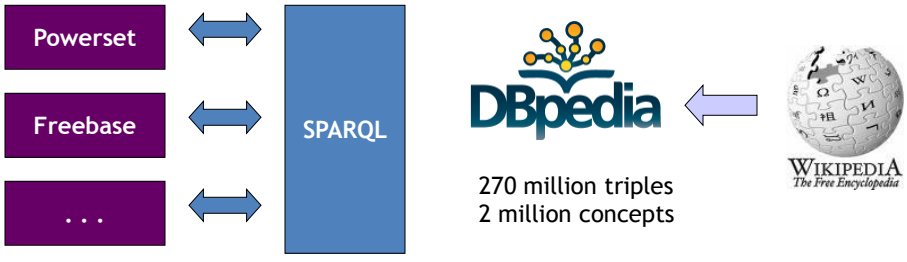
Querying

- The equivalent of SQL is SPARQL

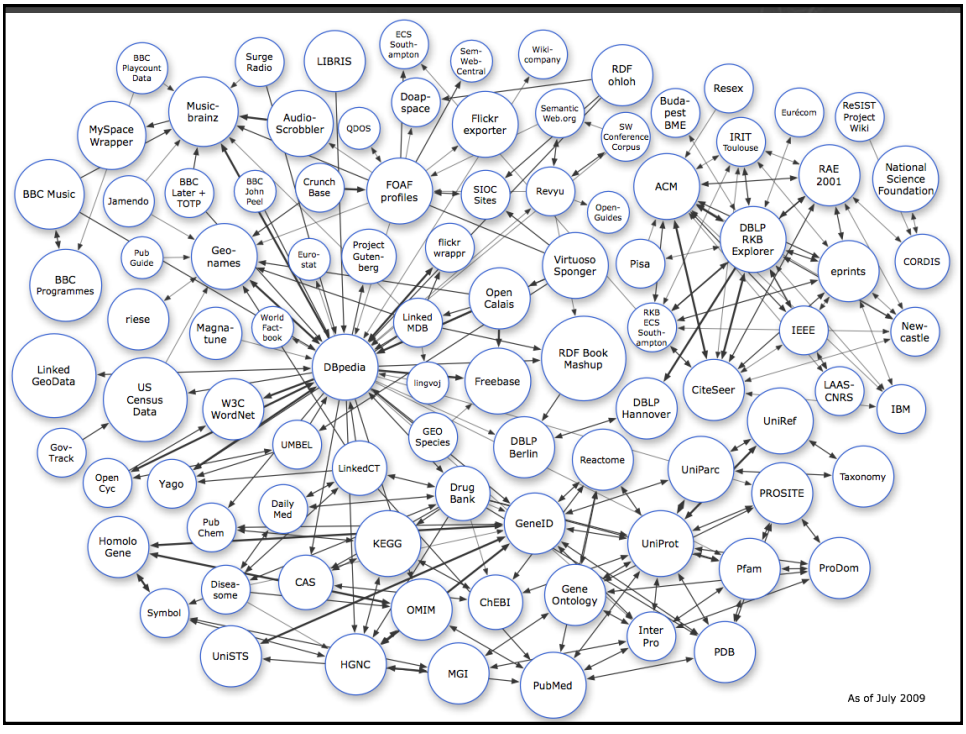
```
PREFIX foaf:
<http://xmlns.com/foaf/0.1/>
SELECT ?name ?mbox
WHERE
   { ?x foaf:name ?name .
     ?x foaf:mbox ?mbox }
```

- A service that accepts SPARQL queries is called a "SPARQL End Point"

SPARQL End Points



270 million triples
2 million concepts



ROA

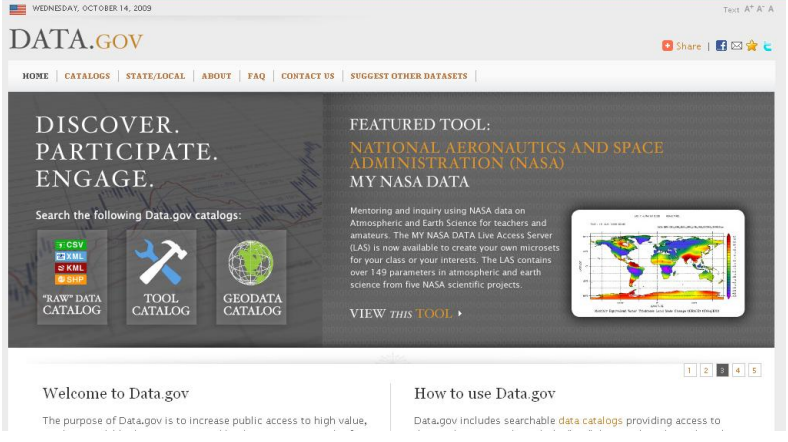
- Semantic Web Technology is a natural fit with Resource Oriented Architecture
  - Based on RESTful design principles
  - Few verbs, many nouns
  - Nouns are the Resources, expressed as URIs

"Web of Things"

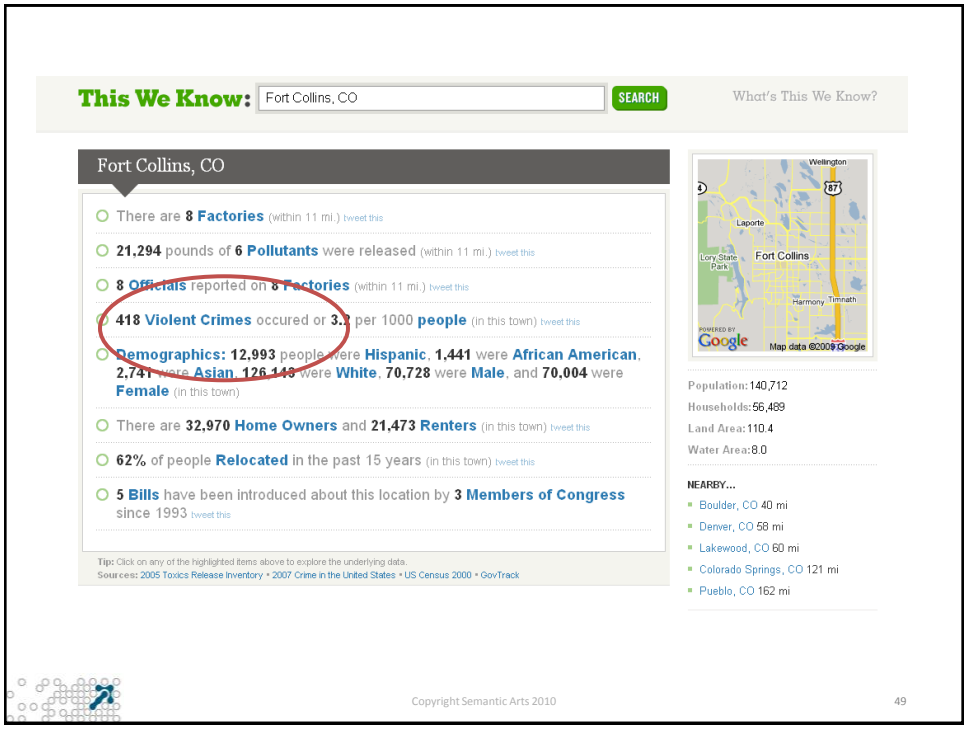
- Semantic Technology is also a good fit for the newly emerging "web of things"
  - Sensors
  - Motes
  - RFID

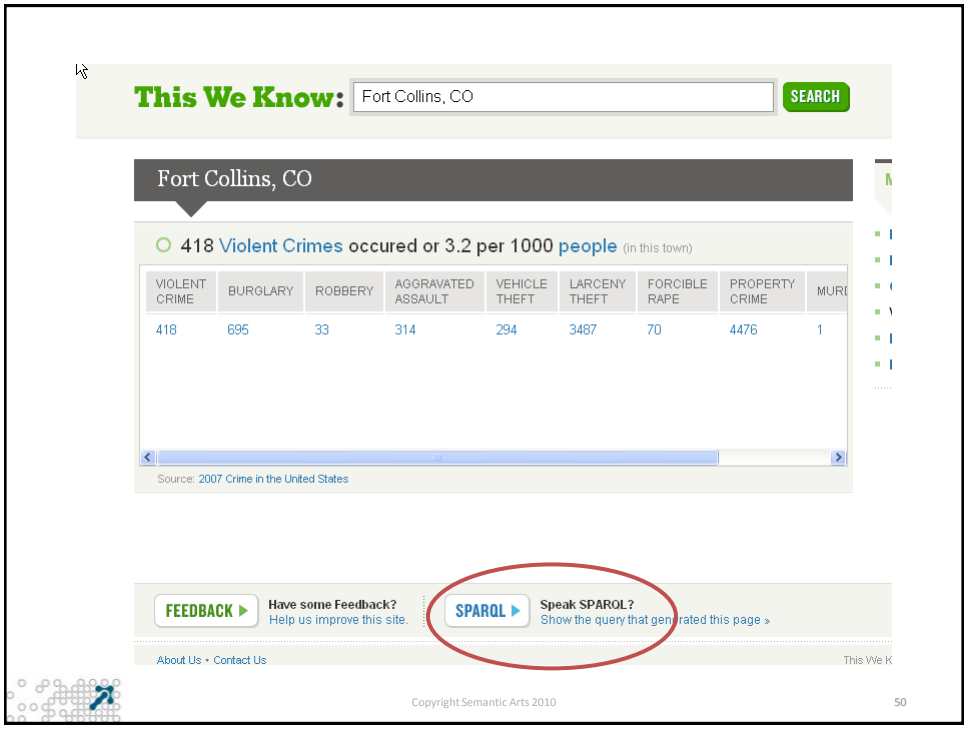
Transparency

- What governments, and companies, are discovering.
- Our chance of becoming wise (acquiring wisdom) will be increased if we can give more people access to more information and knowledge.



http://data.gov





SPARQL ▶ Speak SPARQL?
Show the query that generated this page »

```
  PREFIX o: <http://www.data.gov/ontology#>
PREFIX ui: <http://www.thisweknow.org/ui#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX samp: <http://www.rdfabout.com/rdf/schema/uscensus/details/samp/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

  SELECT ?violent_crime ?town
       ?burglary ?town
       ?robbery ?town
       ?aggravated_assault ?town
       ?vehicle_theft ?town
       ?larceny_theft ?town
       ?forcible_rape ?town
       ?property_crime ?town
       ?murder ?town
       ?population ?town
FROM <data> WHERE {
?crime_stat o:town ?town;
            o:violentCrime ?violent_crime;
            o:burglary ?burglary;
            o:robbery ?robbery;
            o:aggravatedAssault ?aggravated_assault;
            o:motorVehicleTheft ?vehicle_theft;
            o:larcenyTheft ?larceny_theft;
            o:forcibleRape ?forcible_rape;
            o:propertyCrime ?property_crime;
            o:murder ?murder;
            o:population ?population .
FILTER(?town=<http://www.rdfabout.com/rdf/usgov/geo/us/co/counties/larimer_county/fort_collins>)
}
```


Summary

- We are transitioning into the Post-Relational Age
- Our need to transition is that we can't deal with the complexity or rigidity of schema-first design
- Semantic Technology offers a powerful new approach: modular, roll-your-own schema accessing far more data than you will ever have in your corporation
- Compatible with ROA & "Web of Things"

Questions?