Gary King

Querying Federated Knowledge for Web 3.0

Using AllegroGraph to Bring Federation to the Enterprise to help scale and manage Semantic Web data.

meta!

Outline

- Why do we need a semantic web?
- Why do we need Federation?
- The Data opportunity
- Federation scales for big data
- Federation Manages
- Federation Warehouses
- AllegroGraph Federation
- Federation Demonstration

Outline

- Why do we need a semantic web?
- Why do we need Federation?
- The Data opportunity
- Federation scales for big data
- Federation Manages
- Federation Warehouses
- AllegroGraph Federation
- Federation Demonstration

Why do we need a Semantic Web?

- Use distributed knowledge to create new types of analytics and build new services
  - Combine many small knowledge stores into new services: Mash-ups, Facebook Applets
  - Combine large knowledge stores to get data where it needs to be and speed knowledge sharing in the Enterprise.

Outline

- Why do we need a semantic web?
- Why do we need Federation?
- The Data opportunity
- Federation scales for big data
- Federation Manages
- Federation Warehouses
- AllegroGraph Federation
- Federation Demonstration

Why do we need to Federate?

- Data, data everywhere
  - And it comes in all different shapes, kinds and colors
- Scaling and Managing
  - Accessing and loading
  - Tracking and understanding
  - Archiving and warehousing

Federation in AllegroGraph

- Small pieces, loosely joined
  - David Weinberger
- (near) Transparent Data integration for the Enterprise

- *Not semantic mediation*
- *More than SPARQL endpoints*

Federation | fedə rā sh ən|
an organization or group within which smaller divisions have some degree of internal autonomy

meta!

AllegroGraph Federation Allows

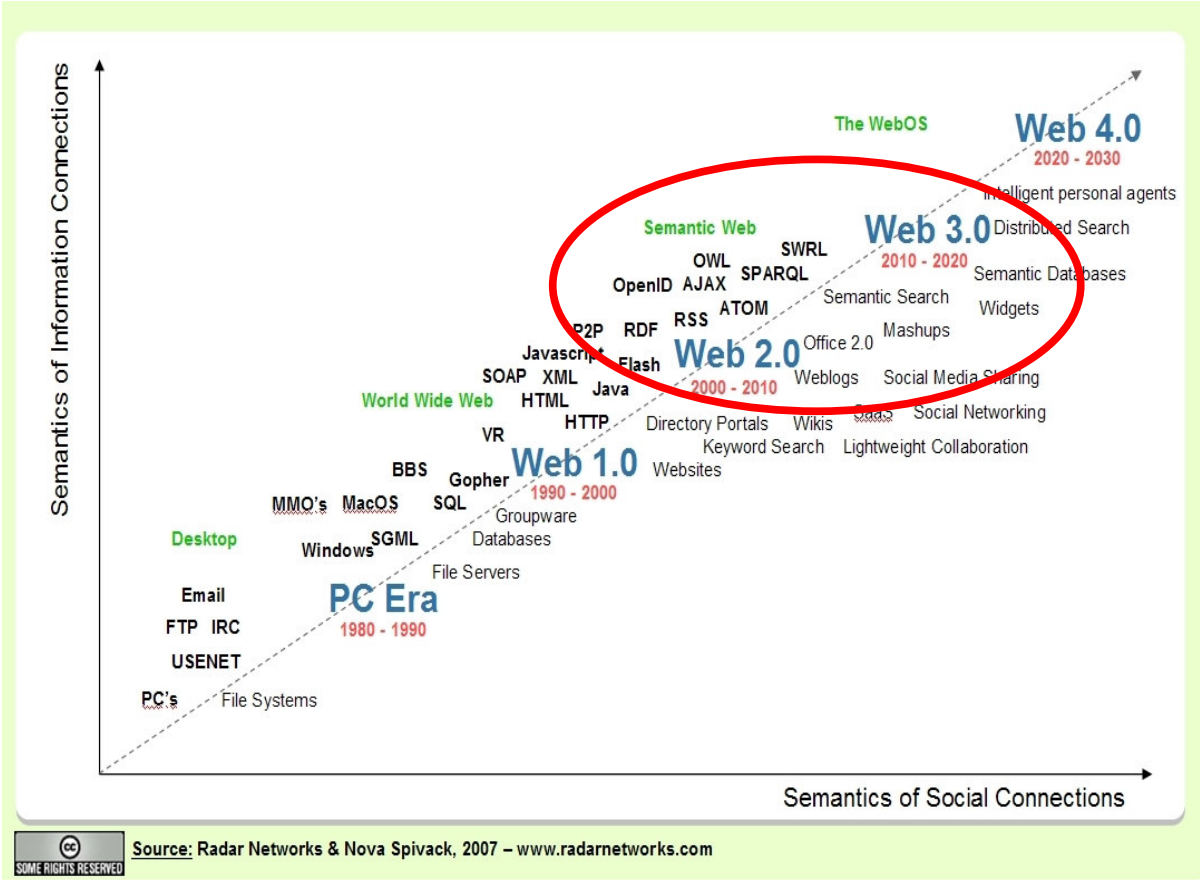
- Building loosely coupled RDF data stores dynamically so that we can
  - Manage data ingest more easily
  - Manage data sources and kinds more flexibly
  - Manage data volume over time
- Because the Federated Triple-store implements all of the AllegroGraph interfaces, we can use all of our usual tools and tricks: Social Network Analysis, Geospatial reasoning, fast range queries, SPARQL, Prolog, and so on!

Outline

- Why do we need a semantic web?
- Why do we need Federation?
- The Data opportunity
- Federation scales for big data
- Federation Manages
- Federation Warehouses
- AllegroGraph Federation
- Federation Demonstration

Data Growth – vive la triple!

- Semantic Web is the next big thing



Data Growth

- More and more knowledge encoded in RDF or transduced into RDF
- More triple-stores in more places
- More triples in every triple-store

A disk manufacturer's dream come true!

meta!

Outline

- Why do we need a semantic web?
- Why do we need Federation?
- The Data opportunity
- Federation scales for big data
- Federation Manages
- Federation Warehouses
- AllegroGraph Federation
- Federation Demonstration

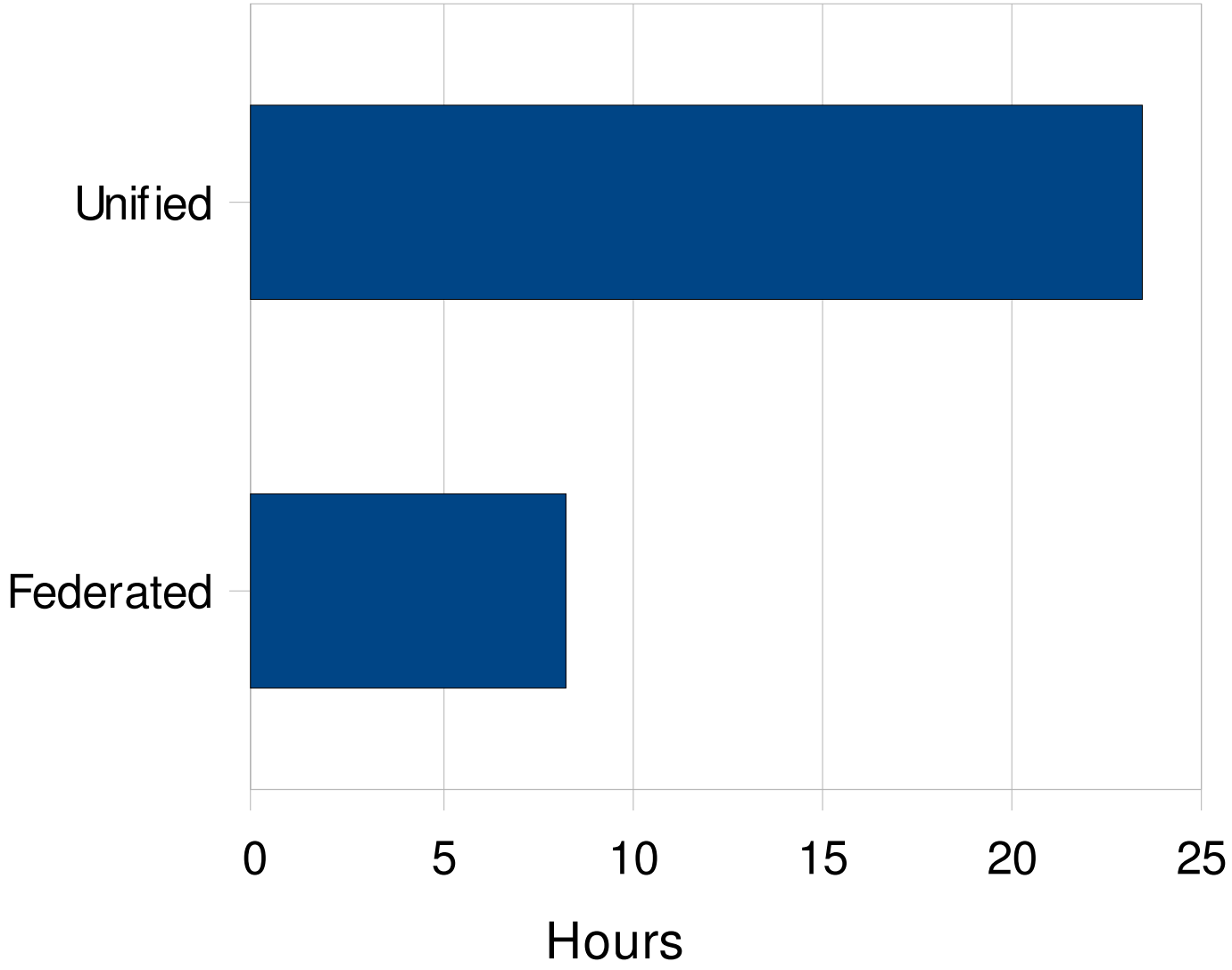
Why Federate – Data ingest

- LUBM is the Lehigh University Benchmark. It models an artificial university with students, departments, professors, classes and all the rest. Though artificial (and highly regular in its structure), it makes a nice test bed for comparing RDF engines and reasoning tools.
- LUBM 8000 is an sample data set modeling 8000 schools. It has more than a billion triples.

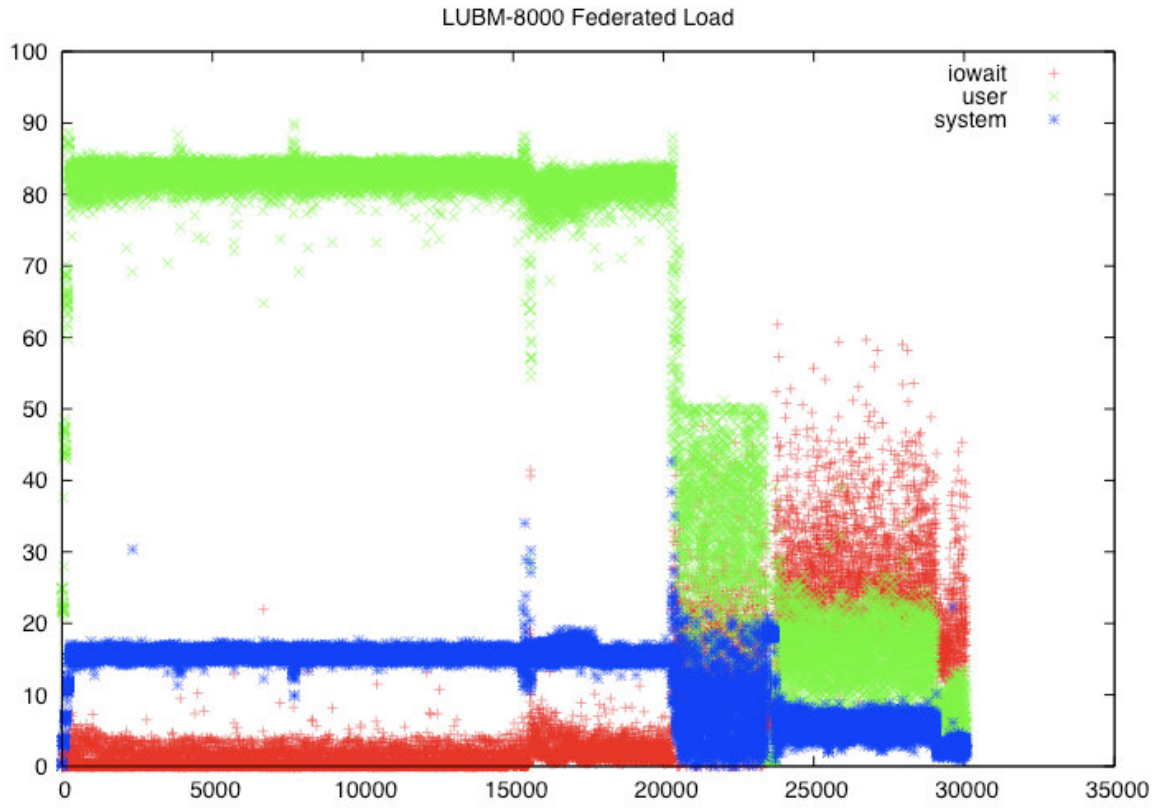
Loading LUBM

- Two scenarios
  - One: a single machine with four CPUs, multiple disks and lots of memory loads and indexes the entire LUBM 8000 dataset by itself
  - Two: the same machine splits the dataset into four triple-stores and then federates these into a single virtual store

Time to Build LUBM 8000



CPU and Disk contention



meta!

LUBM 8000 Demonstration

- Compare query speed for a Federated triple-store (comprised of four roughly equal parts) and a single store.
  - The Federated store lives on two disks
  - The Unified store lives only on one
- The Query

```
(select-distinct (?x)
  (q ?x !ub:takesCourse !u0d0:GraduateCourse0)
  (q ?x !rdf:type !ub:GraduateStudent))
```

Find me the graduate students who take course0 at university 0 in department 0.

meta!

LUBM 8000 Demonstration

```
(select-distinct (?x)
  (q ?x !ub:takesCourse !u0d0:GraduateCourse0)
  (q ?x !rdf:type !ub:GraduateStudent))
```

- Unified store :
  - First run 1.375 seconds
  - Second run 0.002
- Federated store
  - First run 3.489
  - Second run 0.004

meta!

Federation scales

- Federation scales because each of N CPUs can take on roughly 1/N of the work (there is very little contention)
- Federation lets us combine the N triple-stores dynamically when we need to and also lets us spread the work of querying over N CPUs.
- Sorting algorithms grow logarithmically (order N log N), but reducing the number of triples in each store and handling them in parallel, this growth is mitigated and easily managed.
- Load a billion triples over lunch (and a short nap!)

Outline

- Why do we need a semantic web?
- Why do we need Federation?
- The Data opportunity
- Federation scales for big data
- Federation Manages
- Federation Warehouses
- AllegroGraph Federation
- Federation Demonstration

Federation Manages

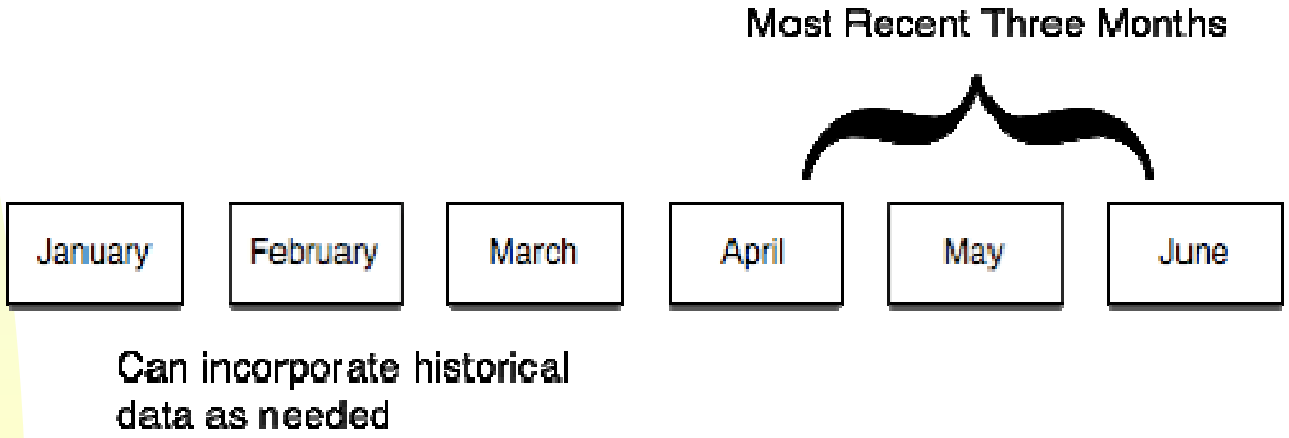
- Real data comes in many varieties
  - Known facts
  - Inferred triples
  - Provenance
  - Ontologies
  - Metadata
  - Deleted triples
- Federation lets us place these data in separate stores: changing ontologies is a snap
- Query optimizers can learn which stores contain which triples

Outline

- Why do we need a semantic web?
- Why do we need Federation?
- The Data opportunity
- Federation scales for big data
- Federation Manages
- Federation Warehouses
- AllegroGraph Federation
- Federation Demonstration

Federation Warehouses

- Enterprise triple-stores need to grow indefinitely and handle vast data volumes over time.
- Federation makes it easy to group triples by date and shift the focus of interest over time

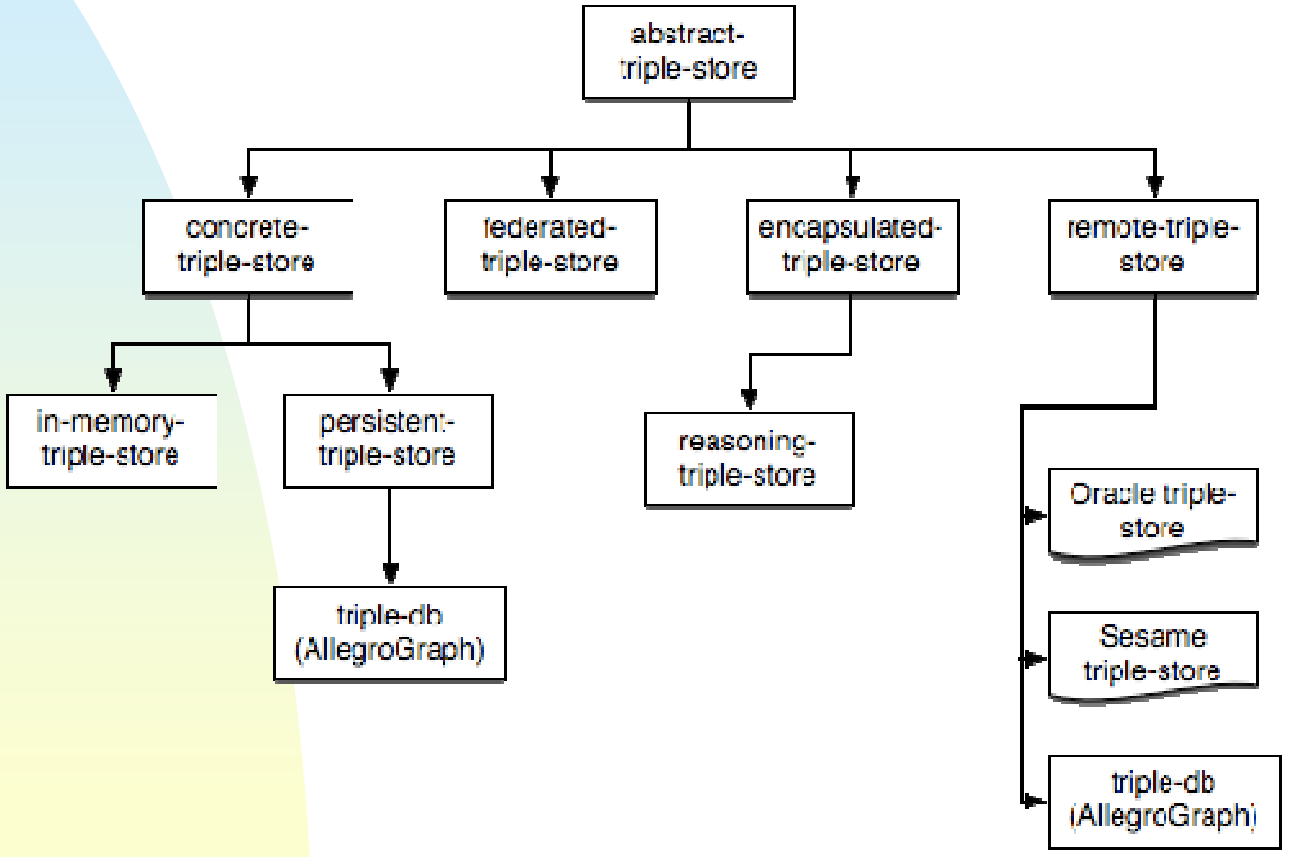


meta!

Outline

- Why do we need a semantic web?
- Why do we need Federation?
- The Data opportunity
- Federation scales for big data
- Federation Manages
- Federation Warehouses
- AllegroGraph Federation
- Federation Demonstration

What can be federated?



meta!

Outline

- Why do we need a semantic web?
- Why do we need Federation?
- The Data opportunity
- Federation scales for big data
- Federation Manages
- Federation Warehouses
- AllegroGraph Federation
- Federation Demonstration

Three (not so) easy pieces

- The *dbPedia lets us learn about famous people*
- *The geonames triple-store lets us geo-code and find places near other places*
- *The US Census data has all manner of information about specific geographic regions*

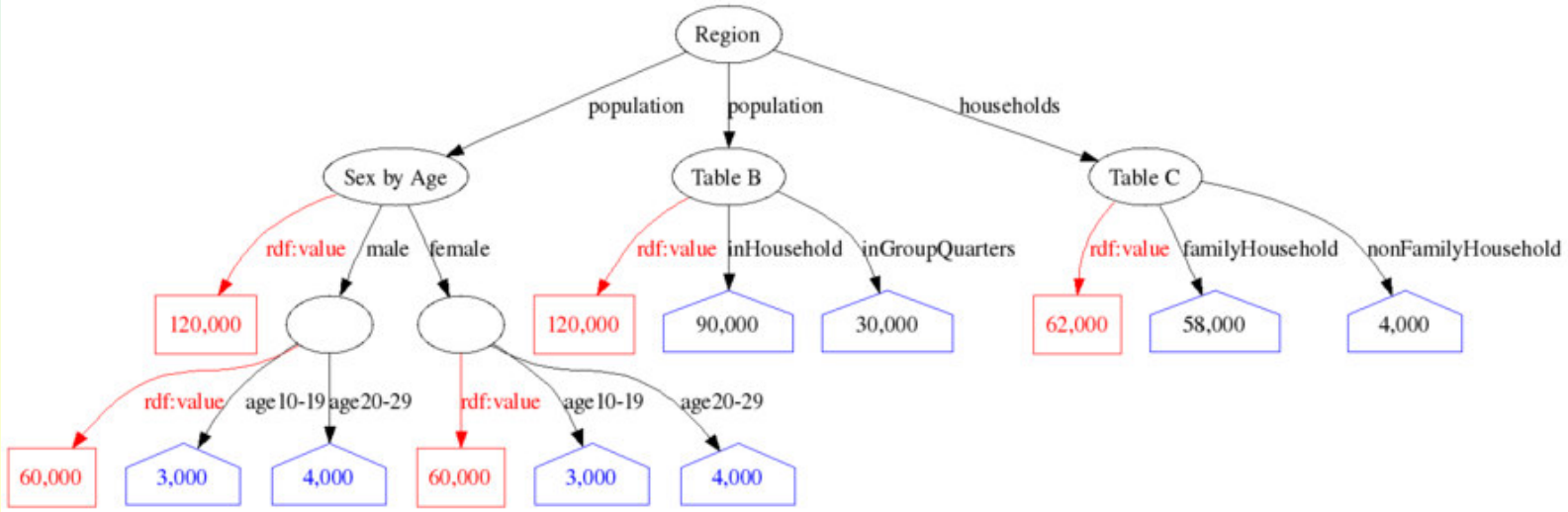
dbPedia

- **DBpedia** is a community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to ask sophisticated queries against Wikipedia and to link other datasets on the Web to Wikipedia data.
  - http://dbpedia.org/About
  - Freie Universität Berlin (and others)

- 218-million triples

Census

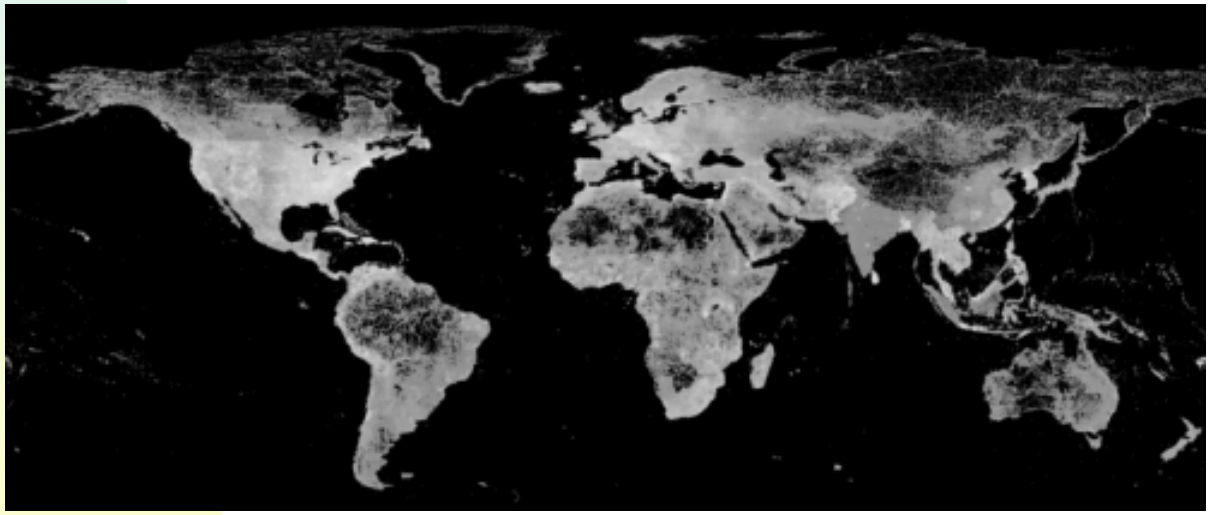
- The U.S. Census data is provided by the Census Bureau in a structured format and yields on the order of 1 billion RDF triples.
  - http://www.rdfabout.com/demo/census/
  - Joshua Tauberer



meta!

Geonames

- The GeoNames geographical database contains over eight million geographical names and consists of 6.5 million unique features whereof 2.2 million populated places and 1.8 million alternate names.
  - http://www.geonames.org/about.html



Light represents information density in the triple-store

What can we ask?

- Where was Barack Obama born?
  - dbPedia
- What cities and towns are within 10 miles of his birthplace?
  - Geonames
- What is the average poverty level of these places in the year 2000?
  - Census

Demonstration

AllegroGraph Federation

- With Federation
  - Load a billion triples over lunch
  - Manage data flexibly
  - Manage data over time
  - Use it transparently
- For more information or follow-up questions, contact Steve Sears (ssears@franz.com)