Graph Databases and the Future of Large-Scale Knowledge Management

Marko A. Rodriguez
T-5, Center for Nonlinear Studies
Los Alamos National Laboratory
http://markorodriguez.com

April 8, 2009

Los Alamos
NATIONAL LABORATORY
EST. 1943

Abstract

Modern day open source and commercial graph databases can store on the order of 1 billion relationships with some databases reaching the 10 billion mark. These developments are making the graph database practical for applications that require large-scale knowledge structures. Moreover, with the Web of Data standards set forth by the Linked Data community, it is possible to interlink graph databases across the web into a giant global knowledge structure. This talk will discuss graph databases, their underlying data model, their querying mechanisms, and the benefits of the graph data structure for modeling and analysis.

Outline

- The Relational Database vs. the Graph Database
- The World Wide Web vs. the Web of Data

Outline

- **The Relational Database vs. the Graph Database**
- The World Wide Web vs. the Web of Data

Our Make Believe World

- Marko is a human and Fluffy is a dog.
- Marko and Fluffy are good friends.

Our World Modeled in a Relational Database

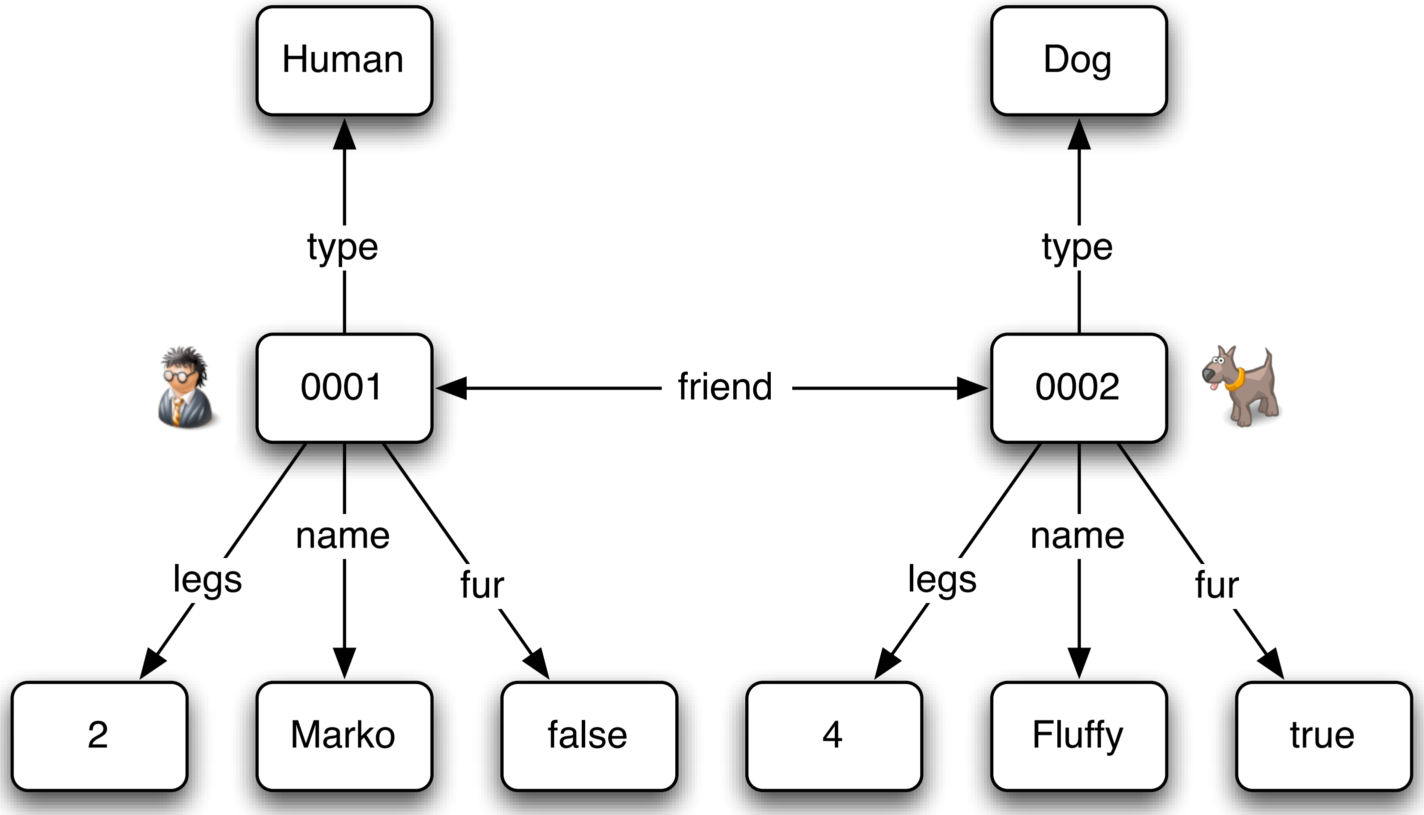
ID	Name	Type	Legs	Fur
0001	Marko	Human	2	false
0002	Fluffy	Dog	4	true

Object_Table

ID2	ID2
0001	0002
0002	0001

Friendship_Table

Our World Modeled in a Graph Database



Extending our Make Believe World

- Marko is a human and Fluffy is a dog.
- Marko and Fluffy are good friends.
- **Human and dog are a subclass of mammal.**

Our Extended World Modeled in a Relational Database

ID	Name	Type	Legs	Fur
0001	Marko	Human	2	false
0002	Fluffy	Dog	4	true

Object_Table

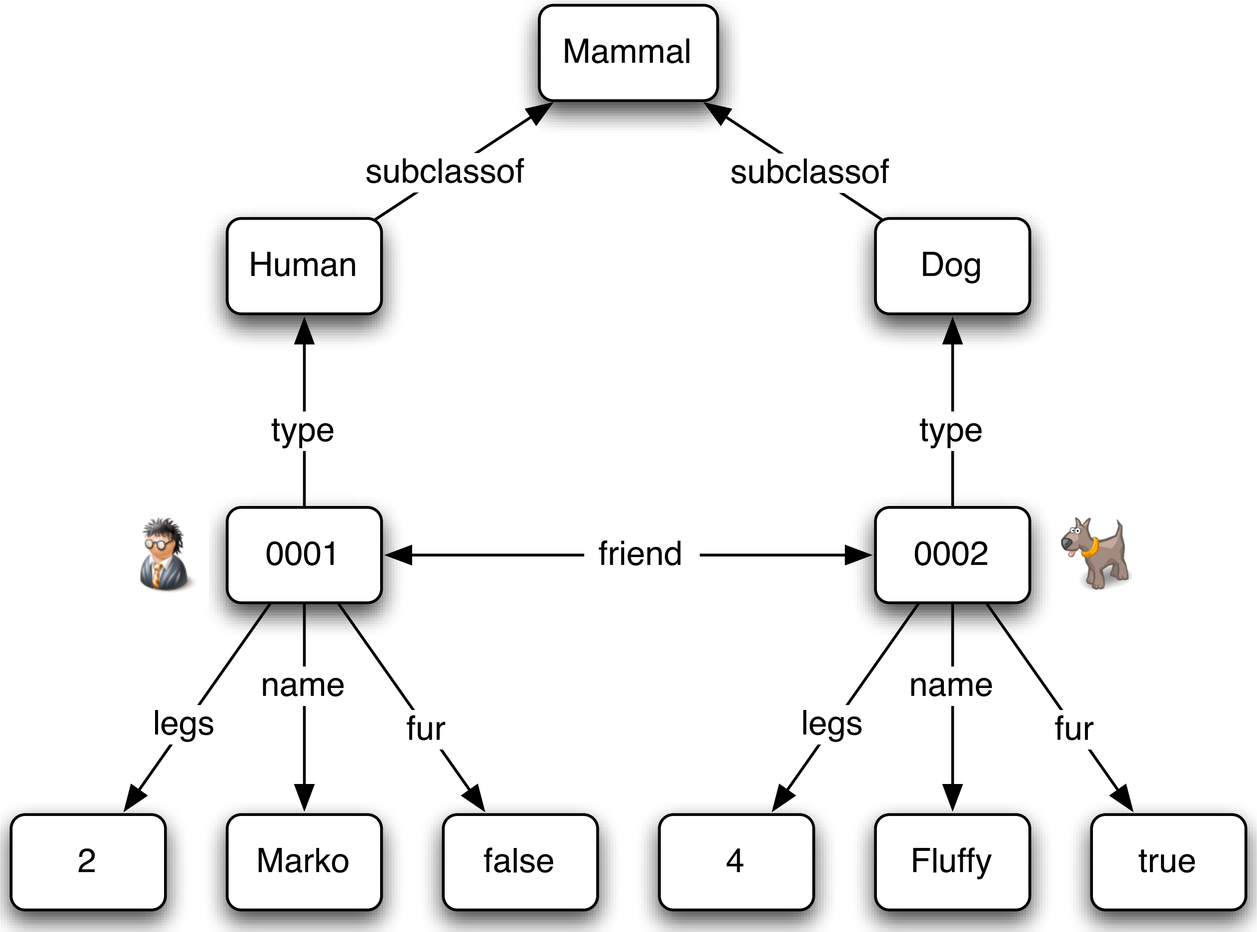
ID2	ID2
0001	0002
0002	0001

Friendship_Table

Type1	Type2
Human	Mammal
Dog	Mammal

Subclass_Table

Our Extended World Modeled in a Graph Database



Extending our Extended Make Believe World

- Marko is a human and Fluffy is a dog.
- Marko and Fluffy are good friends.
- Human and dog are a subclass of mammal.
- **Fluffy peed on the carpet.**

Our Extended Extended World Modeled in a Relational Database

ID	Name	Type	Legs	Fur
0001	Marko	Human	2	false
0002	Fluffy	Dog	4	true
0003	My_Rug	Carpet	N/A	N/A

Object_Table

ID2	ID2
0001	0002
0002	0001

Friendship_Table

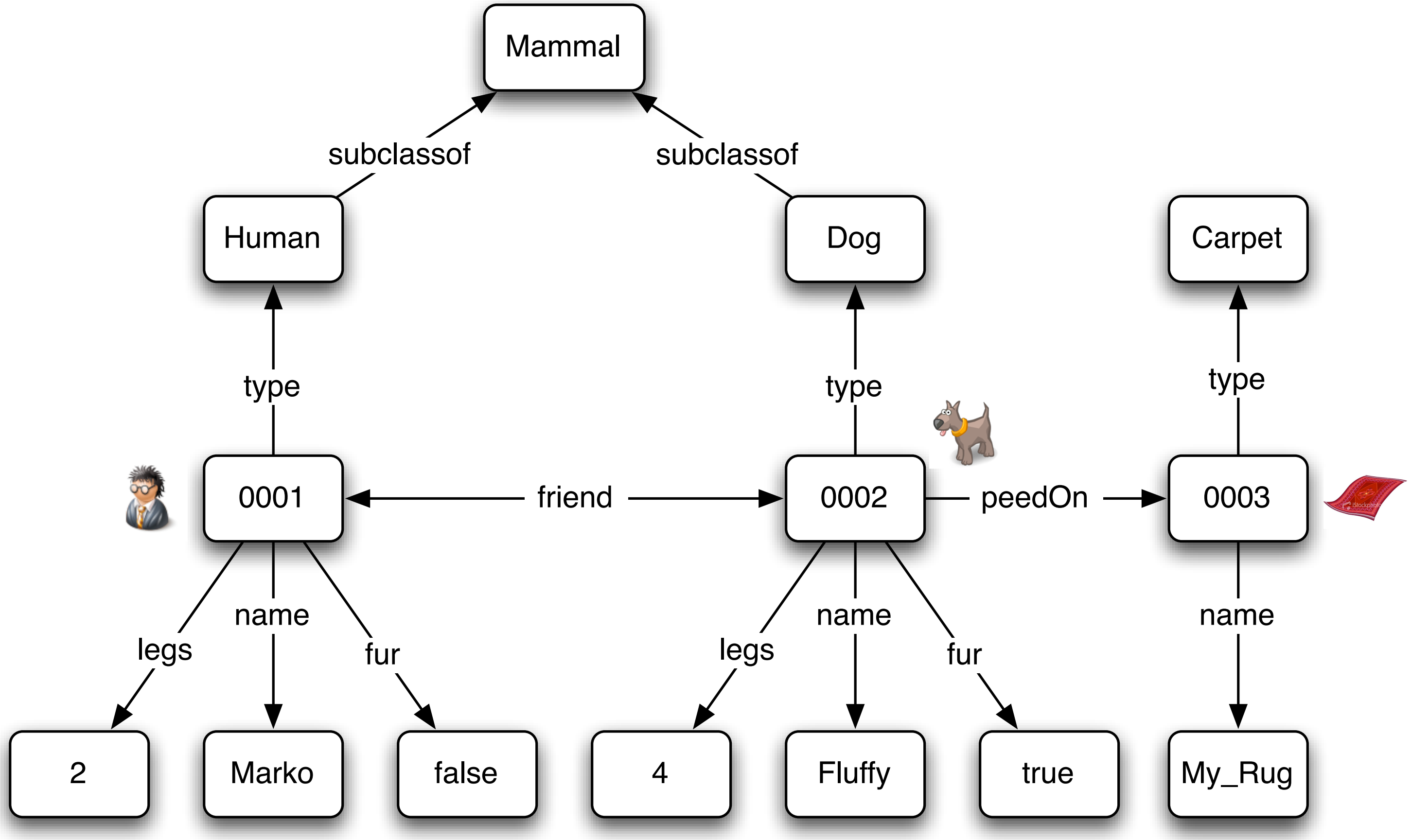
Type1	Type2
Human	Mammal
Dog	Mammal

Subclass_Table

ID1	ID2
0002	0003

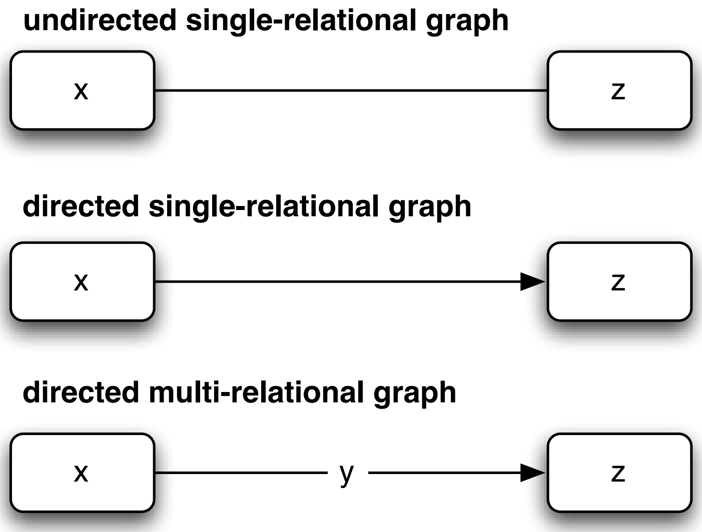
Pee_Table

Our Extended Extended World Modeled in a Graph Database



The Graph as the Natural World Model

- The world is inherently (or perceived as) object-oriented.
- The world is filled with objects and relations among them.
- The multi-relational graph is a very natural representation of the world.


undirected single-relational graph

directed single-relational graph

directed multi-relational graph

The Graph as the Natural Programming Model

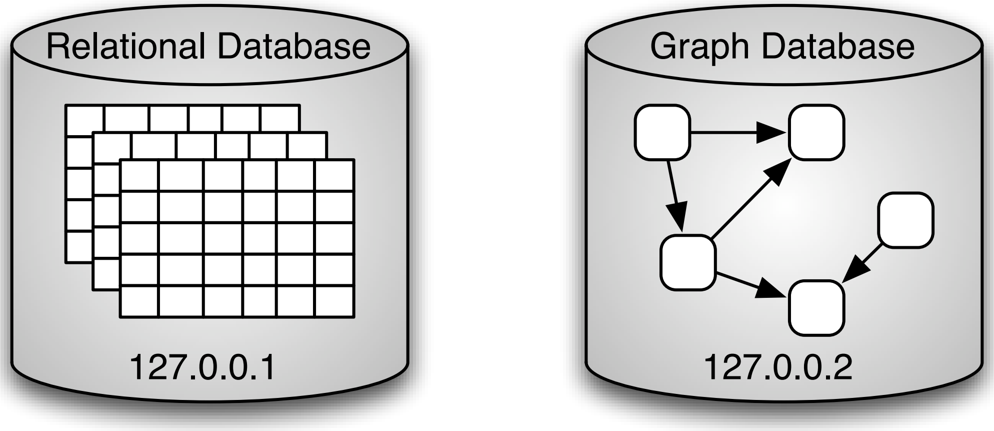
- High-level computer languages are object-oriented.
- Nearly no impedance mismatch between the multi-relational graph and the programming object.
- It is easy to go from graph database to in-memory object.

```
Human marko = new Human();
marko.name = "Marko";
marko.addFriend(fluffy);
marko.setHasFur(false);
marko.setLegs(2);
```

The Relational Database vs. the Graph Database

- A relational database's (e.g. MySQL, PostgreSQL, Oracle) data model is **a collection interlinked tables**.

- A graph database's (e.g. OpenSesame, AllegroGraph, Neo4j) data model is **a multi-relational graph**.



SQL vs. SPARQL

```
SELECT OTY.Name FROM Object_Table AS OTX,
        Object_Table AS OTY, Friendship_Table WHERE
   OTX.Name = "Marko"
   AND Friendship_Table.ID = OTY.ID
   AND Friendship_Table.Friend = OTX.ID;
```

```
SELECT ?z WHERE {
  ?x name "Marko"^^xsd:string .
  ?y friend ?x .
  ?y name ?z }
```

Outline

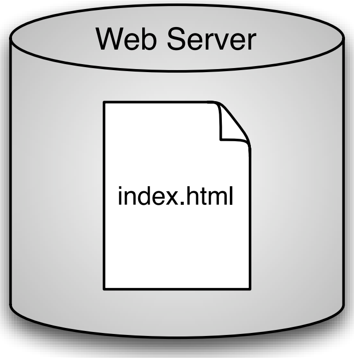
- The Relational Database vs. the Graph Database
- **The World Wide Web vs. the Web of Data**

Internet Address Spaces

- The Uniform Resource Identifier (**URI**) is the superclass of the Uniform Resource Locator (**URL**) and Uniform Resource Name (**URN**).

The Uniform Resource Locator

- The set of all **URL**s is the address space of all resources that can be located and retrieved on the Web. URLs denote **where** a resource is.
  - ⋆ http://markorodriguez.com/index.html
    - ∗ Domain name server (DNS): **markorodriguez.com** → 216.251.43.6
    - ∗ **http://** means GET at port 80,
    - ∗ **/index.html** means the resource to get at that Internet location.



markorodriguez.com
216.251.43.6

The Uniform Resource Name

- The set of all **URN**s is the address space of all resources within the urn: namespace.
    - ⋆ `urn:uuid:bd93def0-8026-11dd-842be54955baa12`
    - ⋆ `urn:issn:0892-3310`
    - ⋆ `urn:doi:10.1016/j.knosys.2008.03.030`

- Named resources need not be retrievable through the Web.

- URNs denote **what** a resource is.

The Uniform Resource Identifier

- The **URI** address space is an infinite space for all Internet resources.
  - ⋆ `http://markorodriguez.com/index.html`
  - ⋆ `urn:issn:0892-3310`
  - ⋆ `ftp://markorodriguez.com/private/markos_secrets.txt`
  - ⋆ `http://www.lanl.gov#fluffy`

- Imporant: URIs can denote **concepts**, **instances**, and **datum**.

lanl:fluffy

lanl:fluffy_legs

The “Uniform Resource Graph”

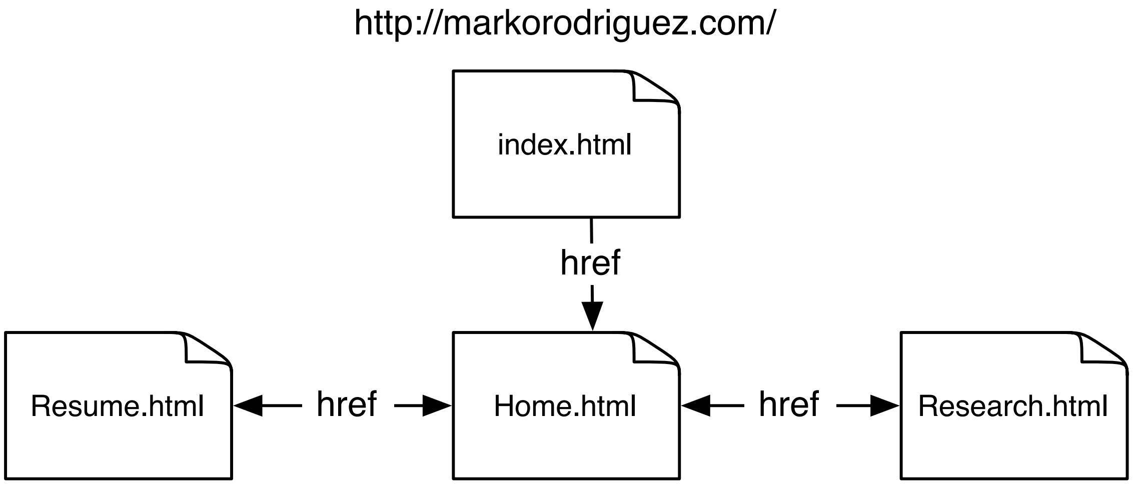
- We can denote **where** something is, **what** something is, but how do we denote how something **relates** to something else?

- How can we denote what something **means**, where meaning is determined by its place within a larger relational structure?
  - ⋆ URIs are like words. They denote things in the real or imaginary world.
  - ⋆ Linking URIs is like defining words. Similar to how a dictionary defines words in terms of other words.

The World Wide Web

- The **World Wide Web** is primarily concerned with the Hyper-Text Transfer Protocol (HTTP) and with retrievable resources in the URL address space.

- These retrievable resources are files: HTML documents, images, audio, etc. The “web” is created when HTML documents contain URLs.



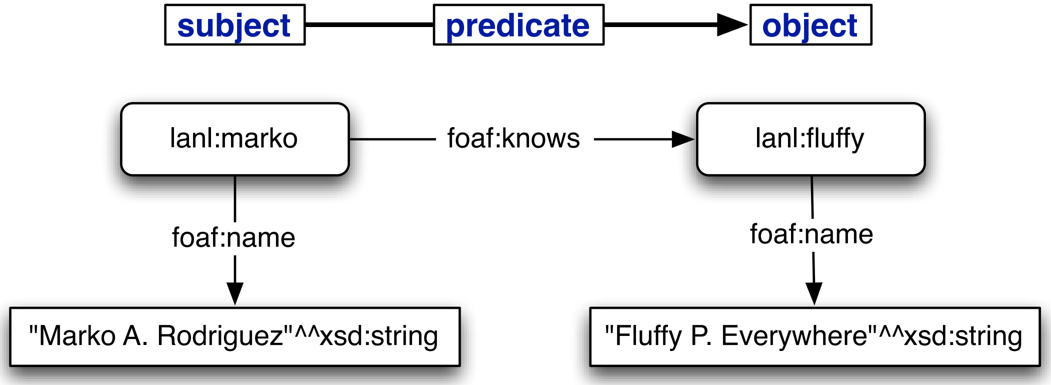
The Web of Data

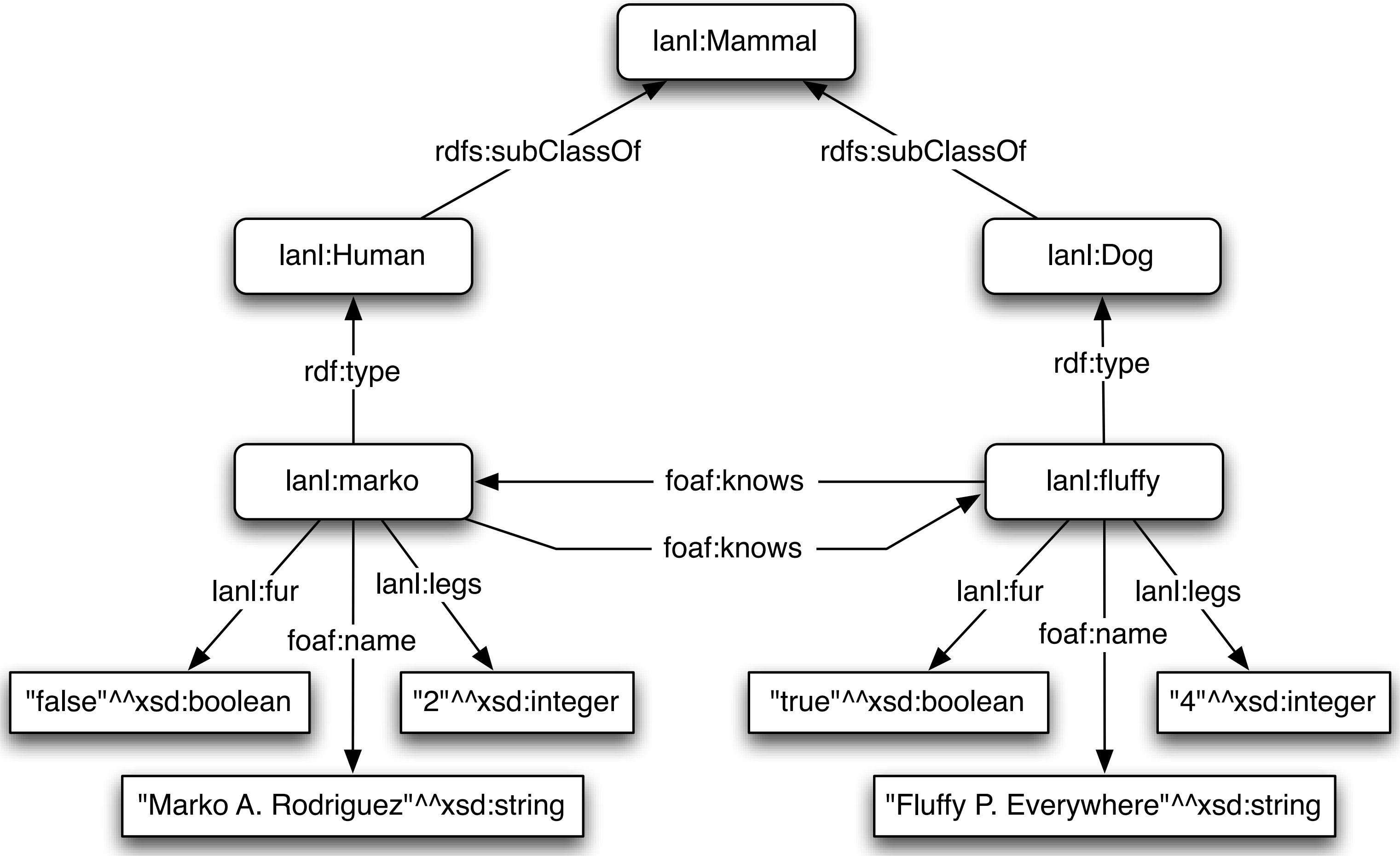
- The **Web of Data** is primarily concerned with URIs. If the World Wide Web is the web of files, the Web of Data is the web of data. In other words, for the World Wide Web, the level of granularity is the retrievable file. For the Web of Data, it is the information in that file. Moreover, this information is not necessarily contained in a file. There existence is predicated on their URI. Their meaning is predicated on their relationship to other URIs. The web of URIs is the Web of Data.

The Resource Description Framework

- The Resource Description Framework (RDF) is the standard for representing the relationship between URIs and literals (e.g. float, string, date time, etc.). I would have preferred the name "Uniform Resource Graph" (**URG**).

- Relationships are directed, labeled links between URIs. A **subject** URI points to an **object** URI or literal by means of a **predicate** URI.





Los Alamos NATIONAL LABORATORY EST. 1943

The RDF Data Model and its Serializations

- RDF is a data model. As such, there exists many serializations (encoding formats) of that model.

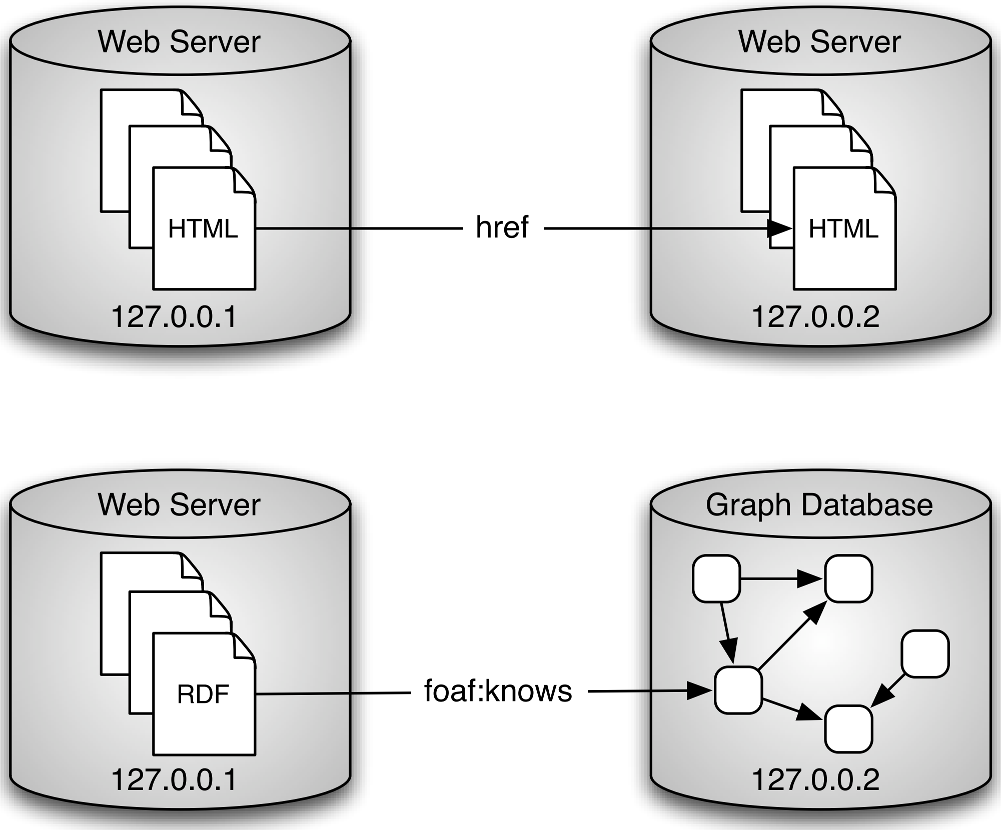
- RDF/XML is **not** RDF. It is a serialization of RDF. It is smart to, at all costs, avoid learning RDF/XML as it is an unintuitive standard. Other serializations include: N-TRIPLE, N3, TRIX, TRIG, ...

```
<http://www.lanl.gov#marko> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.lanl.gov#Human> .
<http://www.lanl.gov#marko> <http://xmlns.com/foaf/0.1/name> "Marko A. Rodriguez"^^<http://www.w3.org/2001/XMLSchema#string> .
<http://www.lanl.gov#marko> <http://www.lanl.gov#legs> "2"^^<http://www.w3.org/2001/XMLSchema#integer> .
<http://www.lanl.gov#marko> <http://www.lanl.gov#fur> "false"^^<http://www.w3.org/2001/XMLSchema#boolean> .
<http://www.lanl.gov#marko> <http://xmlns.com/foaf/0.1/knows>  <http://www.lanl.gov#fluffy> .
```

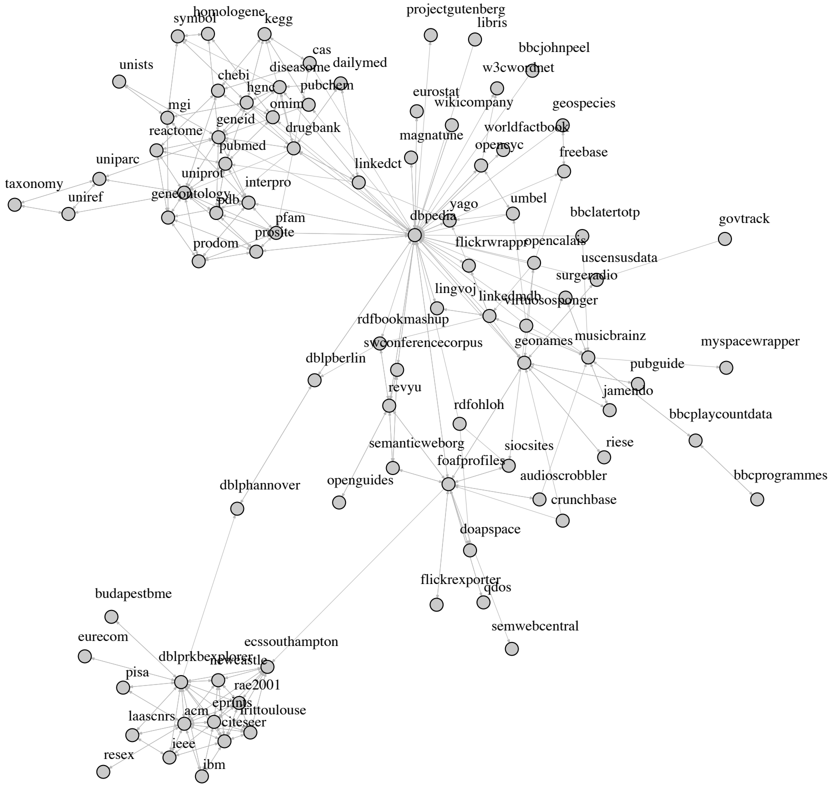
The Web of Data is a Distributed Database

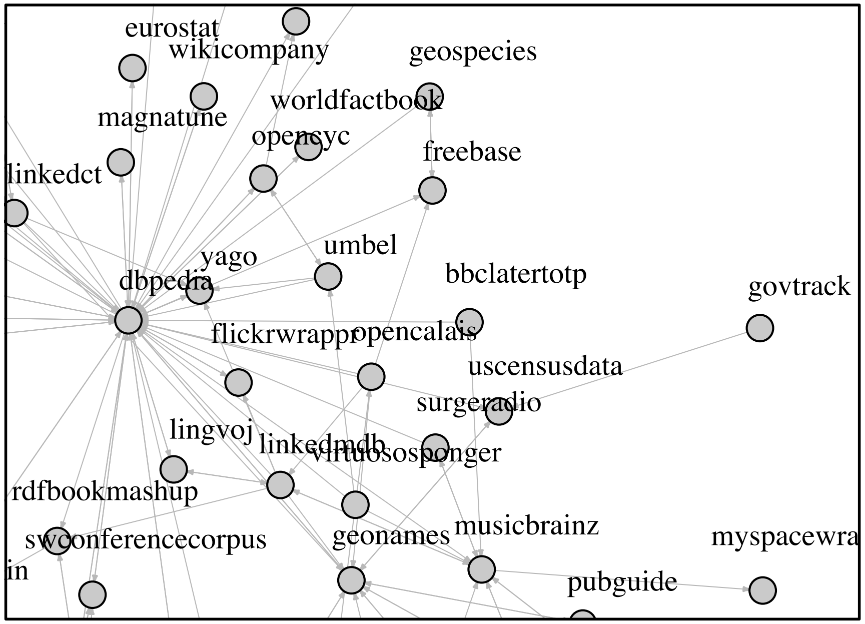
- The URI address space is distributed.
- URIs can denote datum.
- RDF denotes the relationships URIs.
- The Web of Data's foundational standard is RDF.
- Therefore, **the Web of Data is a distributed database**.

The World Wide Web vs. the Web of Data



The Current Web of Data





The Current Web of Data

data set	domain	data set	domain	data set	domain
audioscrobbler	**music**	govtrack	government	pubguide	books
bbclatertotp	music	homologene	biology	qdos	social
bbcplaycountdata	music	ibm	computer	rae2001	computer
bbcprogrammes	media	**ieee**	**computer**	**rdfbookmashup**	**books**
budapestbme	computer	interpro	biology	rdfohloh	social
chebi	biology	jamendo	music	resex	computer
crunchbase	business	laascnrs	computer	riese	government
dailymed	medical	libris	books	semanticweborg	computer
dblpberlin	computer	lingvoj	reference	semwebcentral	social
dblphannover	computer	linkedct	medical	siocsites	social
dblprkbexplorer	computer	linkedmdb	movie	surgeradio	music
dbpedia	**general**	magnatune	music	swconferencecorpus	computer
doapspace	social	musicbrainz	music	taxonomy	reference
drugbank	medical	**myspacewrapper**	**social**	umbel	general
eurecom	computer	opencalais	reference	uniref	biology
eurostat	government	opencyc	general	unists	biology
flickrexporter	images	openguides	reference	uscensusdata	government
flickrwrappr	images	pdb	biology	virtuososponger	reference
foafprofiles	social	pfam	biology	w3cwordnet	reference
freebase	general	pisa	computer	wikicompany	business
geneid	**biology**	prodom	biology	**worldfactbook**	**government**
geneontology	biology	projectgutenberg	books	yago	general
geonames	geographic	prosite	biology	. . .	

Cultural Differences that are Leading to a New World of Large-Scale Knowledge Management

- Relational databases tend to **not** maintain public access points.
- Relational database users tend to **not** publish their schemas.
- Web of Data graph databases maintain public access points called SPARQL end-points.
- Web of Data graph databases tend to reuse and extend public schemas called ontologies.

Conclusion

- Thank you for your time...
  - ⋆ My homepage: http://markorodriguez.com
  - ⋆ Neno/Fhat: http://neno.lanl.gov
  - ⋆ Collective Decision Making Systems: http://cdms.lanl.gov
  - ⋆ Faith in the Algorithm: http://faithinthealgorithm.net
  - ⋆ MESUR: http://www.mesur.org