

# Open Knowledge Network

A.W. Moore & R.V.Guha



# Outline

Ubiquity of Knowledge Bases

The case for an Open Knowledge Network

The essential components of an OKN

Candidate architectures

# Knowledge Graphs are now ubiquitous

Search, Personal Assistants and other consumer apps

We reached the limits of what can be done with text

More form factors and more interaction modalities →  
Structured data is becoming more important ...

Google (KG), Microsoft (Satori), Facebook (OGP), Amazon (Alexa), Apple ...

Each has their own 'knowledge graph'

# Knowledge Graphs in search

 peter gabriel concert dates

**Web** Images Videos Maps News Explore

614,000 RESULTS Any time ▾

**Peter Gabriel Tour Dates 2017, Peter Gabriel Concert ...**  
[www.concertboom.com/peter-gabriel/tickets-2017](http://www.concertboom.com/peter-gabriel/tickets-2017) ▾  
Peter Gabriel Tour Dates. Peter Gabriel tour dates , Peter Gabriel concerts , Peter Gabriel concert ticket . Similar Tours. Radiohead. Coldplay. The Cure. Adele.

**Live Archive - PeterGabriel.com**  
[petergabriel.com/live](http://petergabriel.com/live) ▾  
23 rows · Subscribe to [petergabriel.com](http://petergabriel.com). Enter your email address to receive regular ...

When	Venue	City	Country
21st June 2016	Nationwide Arena	Columbus	United States
23rd June 2016	Verizon Center	Washington	United States

See all 23 rows on [petergabriel.com](http://petergabriel.com)

peter gabriel concert dates

All News Videos Shopping Images More ▾ Search tools

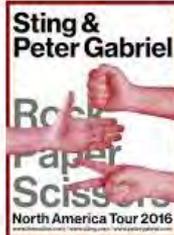
About 417,000 results (0.60 seconds)

Here are the dates:

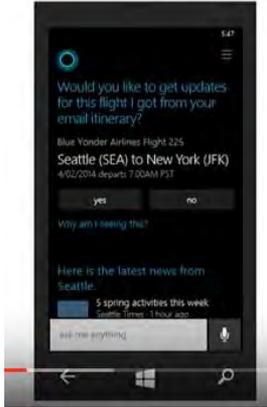
Date	City	Venue
June 21	Columbus, OH	Nationwide Arena
June 23	Washington, DC	Verizon Center
June 24	Wantagh, NY	Jones Beach
June 26	Philadelphia, PA	BB&T Pavilion

17 more rows, 1 more column

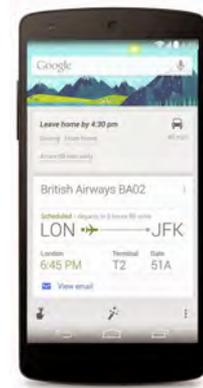
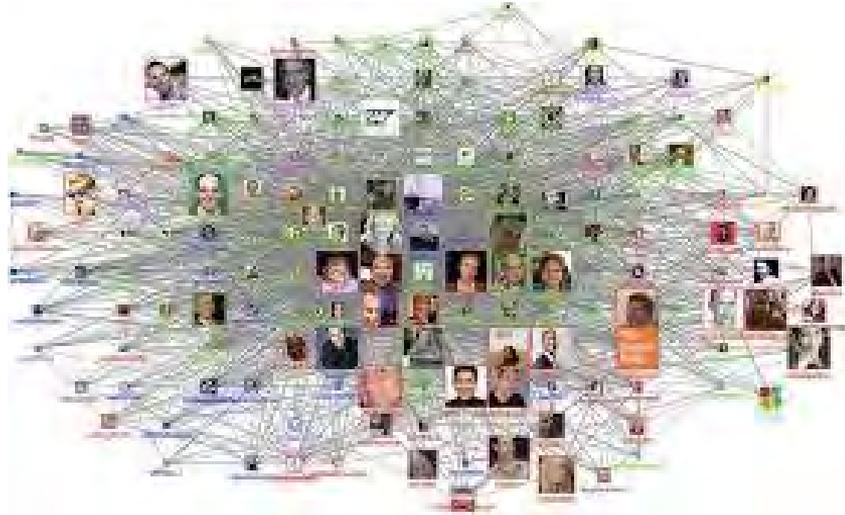
**Peter and Sting Tour 2016 - PeterGabriel.com**  
[petergabriel.com/news/peter-and-sting-tour-2016/](http://petergabriel.com/news/peter-and-sting-tour-2016/) Peter Gabriel ▾



# In Personal assistants



Microsoft  
Cortana



Google  
Now

Google Home &  
Google Assistant  
introduced  
at Google I/O 2016



# Why this initiative: Closed vs Open

Google/Microsoft/... have spent millions to construct these KGs

These KGs are important strategic, closely guarded assets

Hard for broader community to build and extend

Why is this important? Not just a matter of easier access ...

Learning from history: Proprietary Online Services → Web

# Proprietary Online Networks → Web

Online Networks (AOL, Prodigy, ...) from ~1986

High startup costs (dialin systems, content platform, ...)

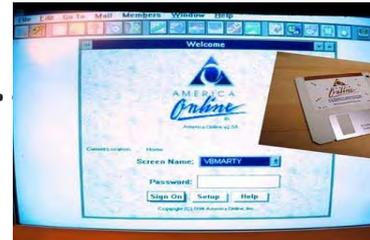
Did not change much over the years

Centralized, comprehensive (payments, identity, ...)

Microsoft entered the field in 1995

All these players were very well funded

Direction decided by small number of people



# Web vs proprietary networks

The web: students in universities, researchers, enthusiasts  
Far from comprehensive: no commerce, no security, ...

But Low upfront costs (leveraged the Internet)  
Don't need anyone's permission try something new  
By December '95, AOL, MSFT kill proprietary services



Why? Remarkable wave of innovation  
Almost every conceivable idea was explored  
Distributed exploration of the design space



# Lessons from history

Knowledge Graphs have demonstrated their utility

But so much more is possible with these:

- Wider range of consumer facing apps

- Much easier integration of data from multiple sources

- Sharing economic/social/scientific data, ...

# Lessons from history

Today Knowledge Graph systems similar to 1986 Online Networks

*Largely isolated, proprietary, monolithic systems whose direction is set by a small number of applications of interest to these companies*

We need to create something more like the web

# Open Knowledge Network

Anybody should be able to add to it

- govt data, scientific data, surf conditions in ...

Any business model for the data

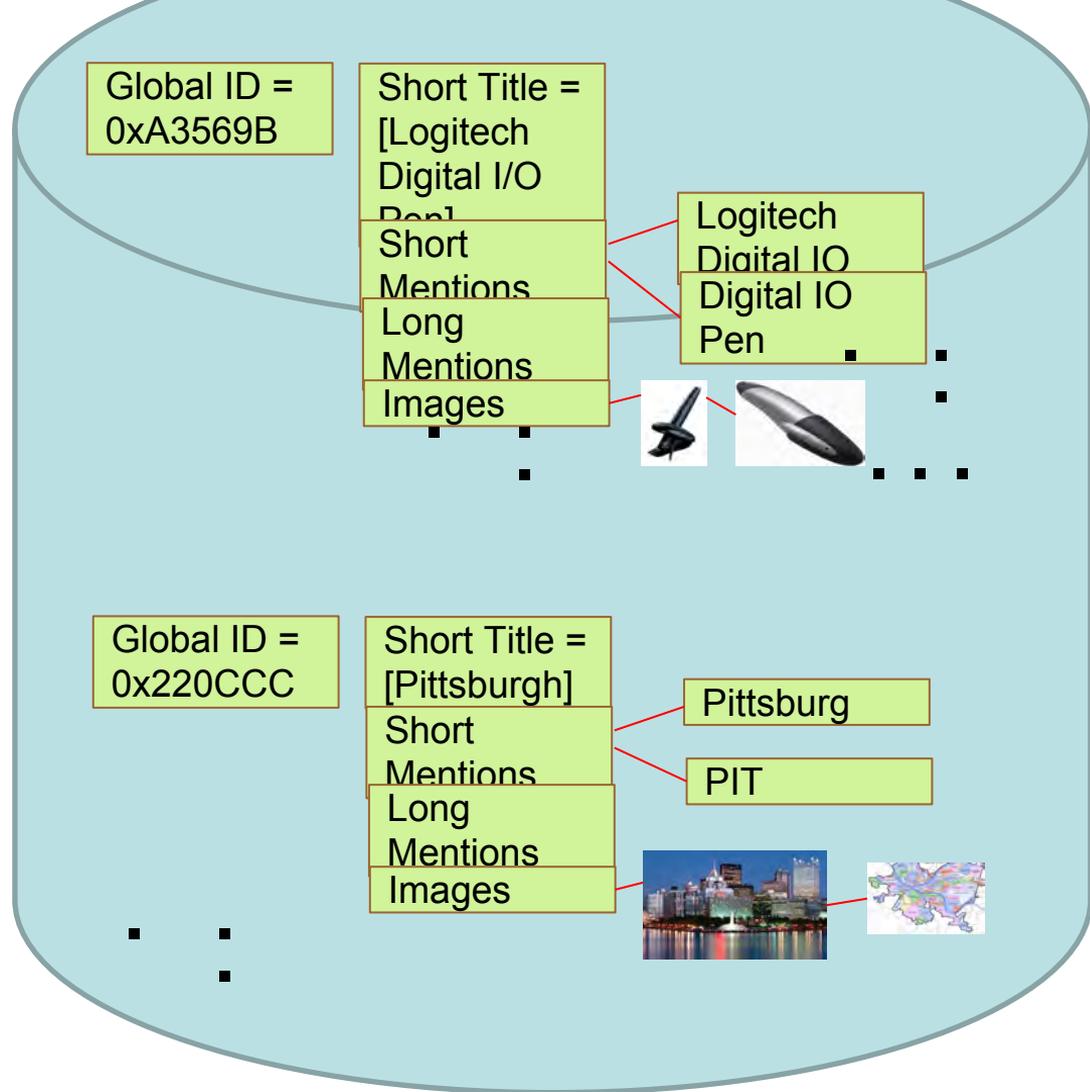
Anyone should be able to build apps using it

**Reminder of the main technical components...**

- Existing Entity Stores
- Architecture
  - Catalog
  - Matching Engine
  - Facts
  - Normalization Engine
- Use Cases
- Risks

- GIS
- Amazon
- UMLS Codes
- Indexed Web Docs
- CYC
- wikidata
- freebase
- tripadvisor etc
- schema.org

- Existing Entity Stores
- Architecture
  - **Catalog**
  - Matching Engine
  - Facts
  - Normalization Engine
- Use Cases
- Risks



- Existing Entity Stores
- Architecture
  - Catalog
  - Matching Engine
  - Facts
  - Normalization Engine
- Use Cases
- Risks

“...unlike **Logi-Techs’ new digital IO stylus**, which...”



Global ID = 0xA3569B with probability 0.94  
Global ID = 0xEEA001 with probability 0.02

- Existing Entity Stores
- Architecture
  - Catalog
  - Matching Engine
  - Facts
  - Normalization Engine
- Use Cases
- Risks

“Triples” is one popular approach:

- <Banana ID>.color = <yellow ID>
- (<HSBC ID> is\_a <Bank ID>)
- (<Dell XPS 13” notebook 2015 ID> has\_a <2mm 12 Volt DC composite power socket ID>)
  - There is and will continue to be a major intellectual war on the expressiveness of the semantics.
  - Winner should be decided by use cases.

- Existing Entity Stores
- Architecture
  - Catalog
  - Matching Engine
  - Facts
  - Normalization Engine
- Use Cases
- Risks



“Jersey”



“NYC”

“Big Apple”



- Existing Stores

- Architect

- Catalog

- Match

- Facts

- Normal

- Engine

- Use Case

- Risks

## Question answering:

### Fact Questions:

- *[How old is vice president Pence?]*
- *[Which Washington-based think tanks have worked on projects involving South American trade?]*
- *[Which building am I in? Where do I go for a taxi?]*

### Research Questions:

- *[What are good things to do with kids in Pittsburgh?]*
- *[Which Hodgkins Lymphoma treatments are covered under the Affordable Care Act for my mother?]*
- *[What do the cells in capillary systems of liver tumors unresponsive to sorenafib have in common?]*

## The right-click on a spreadsheet-column use case

A scientist or analyst wishes to canonicalize and then do joins with data she is using.

## Knowledge-powered machine learning

Allowing secondary and tertiary features and aggregates to be used in machine learning algorithms.

## Knowledge-powered robotics

Common sense reasoning; a robot needs to understand, not simply sense, its environment.

## Knowledge powered startup and app developer ecosystem

Generally making it easier to write a useful app for domain X which needs to know about entities in domain Y (e.g. a great liver cancer app actually needs to know bus routes to treatment centers).

- Existing Entity Stores
- Architecture
  - Catalog
  - Matching Engine
  - Facts
  - Normalization Engine
- Use Cases
- **RISKS**

### Technical Risks

- Undermerging, Overmerging, Multilevel taxonomies, Time, Uncertainty, Provenance.
- **Entity stores are alive:** You don't build an entity store once; you build a process to maintain, grow, and update a set of entities.
- **Physics and Sensing:** Many use cases (robotics and sensing) need to maintain information about visual, acoustic, and physics of physical-world objects.

### Non-technical risks

- **Privacy.** Very serious problem. We recommend not including PII in such a project. There will need to be practical privacy technology in place to ask "what is the average age of women in Pittsburgh?" without having any explicit representation of all the people in Pittsburgh.
- **Provenance:** many major industries have their business model around obtaining facts.
- Why not leave this up to a large internet company to build? (Ans: this is bigger than Google or Apple or....)

# Open Knowledge Network

- Lots of publishers, small to big
- Small set of core protocols and vocabulary
- Services (ala search) that make it easy to build apps
- Web style architecture



# Making progress: Open Data

Directory of datasets

- data.gov: 177,928 data sets
- dataMed: 1,541,00 data sets
- dataverse: 48,112 data sets

Tossing data over the wall ...

One Web vs 100k FTP sites/Million word docs

# Making Progress: Curated Data

## Integrated Vertical Specific Repositories

- Sloan Sky Survey by Jim Gray
- GenBank
- ImageNet

Datasets are driving research breakthroughs

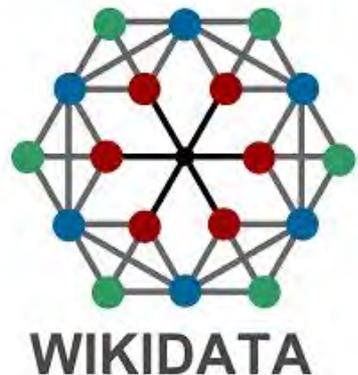
Each of them is very narrow



# Almost there ...

Broad but centrally controlled

- Wikidata



Broad, decentralized, but hard to consume, very limited vocab

- Schema.org

The Schema.org logo consists of a dark red square with the text "schema" in white lowercase letters on the top line and ".org" in white lowercase letters on the bottom line.

# Schema.org

Started in 2011 as a collaboration between Google, Yahoo, MSFT  
Now used by Siri, Google Assistant, etc.

Vocabulary/Schemas for structured data on the web

Web pages, email addresses, ...

Search (structured data in search) was driving application

# 2016 Feeding Academic Success Culinary Challenge and Wellness Expo

Real Food for Kids

Saturday, March 12, 2016 from 10:00 AM to 2:00 PM (EST)

Fairfax, VA



## Ticket Information

TYPE	REMAINING	END		QUANTITY
<b>Culinary Challenge Competition</b> 10:00AM to 11:15PM   Awards 1:30PM to 2:00PM	151 Tickets	Mar 12, 2016	Free	0
<b>Food Is Hot I</b> Session 1 - 11:30 AM to 12:10 PM	9 Tickets	Mar 12, 2016	Free	0
<b>Food Is Hot II</b> Session 2 - 12:20 PM to 1:00 PM	10 Tickets	Mar 12, 2016	Free	0
<b>Food for Thought I</b> Session 1 - 11:30 AM to 12:10 PM	6 Tickets	Mar 12, 2016	Free	0



Save This Event

## When & Where





# Structured Data Testing Tool

http://www.eventbrite.com/e/2016-feeding-i

FETCH & VALIDATE

CANCEL

Shortlink

Results - Filter by use case

```

1 <!DOCTYPE html>
2 <!--[if lt IE 7 ]> <html class="ie ie6"
  xmlns="http://www.w3.org/1999/xhtml"
  xmlns:fb="http://ogp.me/ns/fb#" lang="en-us"> <![endif]-->
3 <!--[if IE 7 ]> <html class="ie ie7"
  xmlns="http://www.w3.org/1999/xhtml"
  xmlns:fb="http://ogp.me/ns/fb#" lang="en-us"> <![endif]-->
4 <!--[if IE 8 ]> <html class="ie ie8"
  xmlns="http://www.w3.org/1999/xhtml"
  xmlns:fb="http://ogp.me/ns/fb#" lang="en-us"> <![endif]-->
5 <!--[if IE 9 ]> <html class="ie ie9"
  xmlns="http://www.w3.org/1999/xhtml"
  xmlns:fb="http://ogp.me/ns/fb#" lang="en-us"> <![endif]-->
6 <!--[if (gt IE 9)|(IE)]><!--> <html
  xmlns="http://www.w3.org/1999/xhtml"
  xmlns:fb="http://ogp.me/ns/fb#" lang="en-us"> <!--<![endif]-->
7 <head>
8 <meta http-equiv="X-UA-Compatible"
  content="IE=edge,chrome=1">
9 <title>2016 Feeding Academic Success Culinary Challenge

```

s. To help illustrate this point, this workshop features a group of spectacular guest stars. \_\_\_\_\_

### offers [Offer]:

<b>name:</b>	Culinary Challenge Competition 10:00AM to 11:15PM I Awards 1:30PM to 2:00PM
<b>url:</b>	http://www.eventbrite.com/e/2016-feeding-academic-success-culinary-challenge-and-wellness-expo-tickets-17823606888
<b>availabilityEnds:</b>	2016-03-12T10:00:00-05:00
<b>price:</b>	0.00
<b>priceCurrency:</b>	USD
<b>inventoryLevel [QuantitativeValue]:</b>	
<b>name:</b>	151 Tickets

### offers [Offer]:

<b>name:</b>	Food Is Hot I Session 1 - 11:30 AM to 12:10 PM
<b>url:</b>	http://www.eventbrite.com/e/2016-feeding-academic-success-culinary-challenge-and-wellness-expo-tickets-17823606888
<b>availabilityEnds:</b>	2016-03-12T12:00:00-05:00
<b>price:</b>	0.00

# Schema.org applications: search



xanh mountain view

About 8,810,000 results (0.26 seconds)

Everything

Pages

Images

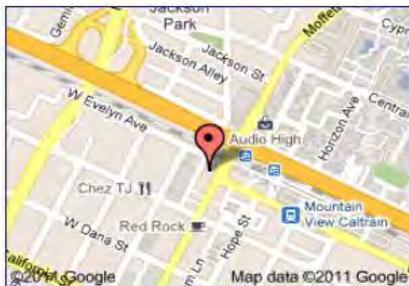
Maps

View, CA  
Location

Tools

**XANH RESTAURANT**

[xanhrestaurant.com/](http://xanhrestaurant.com/) - Cached - Similar



**Xanh Restaurant**   
[Place page](#)

110 Castro Street  
Mountain View, CA 94041  
(650) 964-1888

Tram: Mountain View Station (1)  
[Get directions](#) - [Is this accurate?](#)

Open Mon-Thu 11:30am-2:30pm, 5pm-10pm;  
Fri-Sat 11:30am-2:30pm, 5pm-11pm; Sun  
5pm-10pm

★★★★★ 1587 reviews - [Write a review](#)  
"Pros: Good decor Cheap food ( \$12 lunch)  
Fast service ... Cons: 1. Parking is a ..."  
- [yelp.com](http://yelp.com)

**Xanh Restaurant - Mountain View, CA**

★★★★★ 913 reviews - Price range: \$\$  
913 Reviews of **Xanh Restaurant** "I've been here for dinner and lunch a few times. The dinner is good but I heart their lunch buffets."  
[www.yelp.com](http://www.yelp.com) > Restaurants > Vietnamese - Cached - Similar

**XANH Restaurant - Mountain View, CA** | [OpenTable](#)

★★★★★ 228 reviews - Price range: \$30 and under  
228 Reviews for **XANH** in **Mountain View**. "Nice/fancy presentation but food is so so. The fish is overdone. In general, it is lacking the freshness..."  
[www.opentable.com/xanh](http://www.opentable.com/xanh) - Cached - Similar

**Xanh - Mountain View** | [Urbanspoon](#)

★★★★★ 12 reviews - Price range: \$25 on up per entree  
**Xanh**, Vietnamese Restaurant in **Mountain View**. See the menu, 3 photos, 7 critic reviews, 1 blog post and 4 user reviews. Reviews from critics, food blogs and ...



# Reservations → Personal Assistant

Open Table → confirmation email → Now/Cortana Reminder

 **Cascal Reservations** <member\_services@opentable.com> Jul 22  
to RV ▾

Dear RV,

Thank you for making your reservation through Yelp. You're dining at **Cascal!**  
[Invite your party >](#)

--- Your Reservation Details ---

Diner's name: **RV Guha**  
Date: **Monday, July 22, 2013**  
Time: **8:30 PM**  
Party Size: **2**

[Click here](#) to make changes to your reservation.

**Cascal**  
**400 Castro St. Mountain View, CA 94041**  
**Cross Street: California St.**  
**[\(650\) 940-9500](#)**

[See menus, map & more >](#)

```
<span itemscope
  itemtype="http://schema.org/Restaurant"
  itemid="/restaurant">
  Cascal
</span>

....
<span itemprop="address"
  itemscope
  itemtype="http://schema.org/PostalAddress">
  <span itemprop="streetAddress">
    400 Castro St. Mountain View, CA 94041
  </span>
</span>

...
```

# Schema.org ... the numbers

In use by ~20 million sites: 20% growth over last 12 months

Roughly 35% of pages in search index have markup

~50% of US/EU ecommerce emails

Vocab: Core (~ 2k terms) + extensions (real estate, finance, etc.)

Supported by most major web publishing platforms (Drupal, etc.)

# Schema.org: Major sites

News: Nytimes, guardian, bbc,

Movies: imdb, rottentomatoes, movies.com

Products: ebay, alibaba, sears, cafepress, sulit, fotolia

Local: yelp, allmenus, urbanspoon

Events: wherevent, meetup, zillow, eventful

....

Missing: data.gov, datamed, dataverse, ...

# Schema.org's role

800B small graphs, each with ~25 triples

Would prefer smaller number of bigger graphs

Still too centralized, too focussed on web applications

Data model limited: lacks time, compositionality, ...

Good starting point, but much work needs to be done

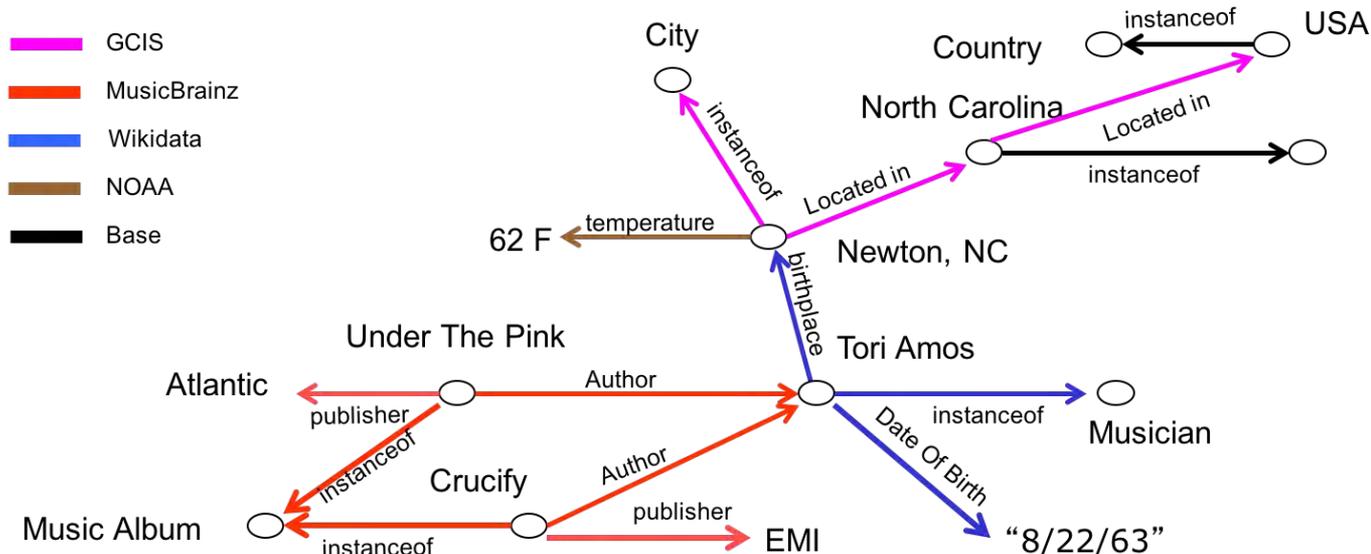
# Lots of interesting problems

- Whats the analog of http/html?
  - How many schemas? 1? 100? 100k?
- Are simple graphs enough?
  - N-ary relations, negations, ... embeddings?
- What's the analog of a search engine
  - stitching together millions of fragments

# Google for data

Google allows user to pretend that the Web is one site

Google for data, for use by programs: Enable developer to pretend all this data is in one database



# Coordinating Names

~1000s of terms like Actor, birthdate

10s for most sites

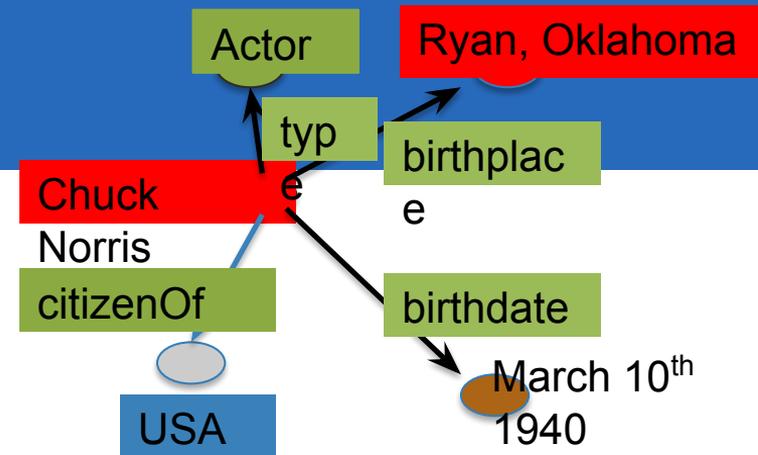
~1b-100b terms like Chuck Norris and Ryan, Oklahoma

Cannot expect 1000s of sites to coordinate on these

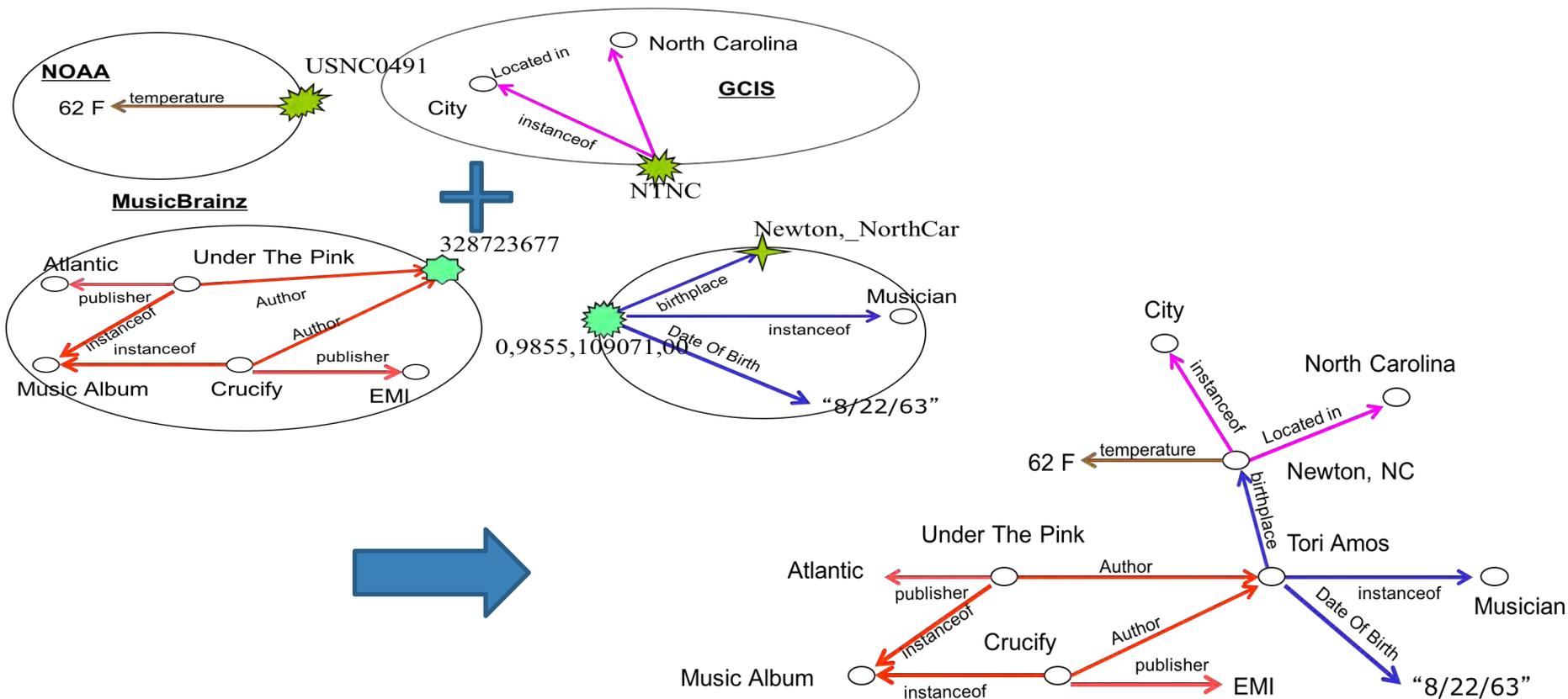
Problem not in generating URIs, Problem in coordination costs

Need to reduce shared vocabulary to minimum!

Agreements  $O(\#cols)$  not  $O(\#rows)$



# The Game of the Name



# Concluding

- Many interesting research problems
- Good news:
  - Don't have to solve all these problems before useful things can be done
  - Lot of data already being published
  - First generation of apps already there

Questions?

**Reserve slides below (may not be needed)**

## Where we are now

More than 15m sites publishing snippets of structured data using schema.org, Facebook OGP, etc.

Biggest problem --- getting publishers to publish is starting to see solution

Search & personal assistants are killer app

# New York Talk Events

RELEVANCE

DATE



MON, MAY 2 6:30 PM

## Theater Talks: Turn Me Loose

Schomburg Center for Research in Black Culture

FREE

#filmmedia #seminar



FRI, APR 22 8:00 PM

## Tribeca Talks After the Movie - Special Correspondents

John Zuccotti Theater @ BMCC Tribeca Performing Arts Center)

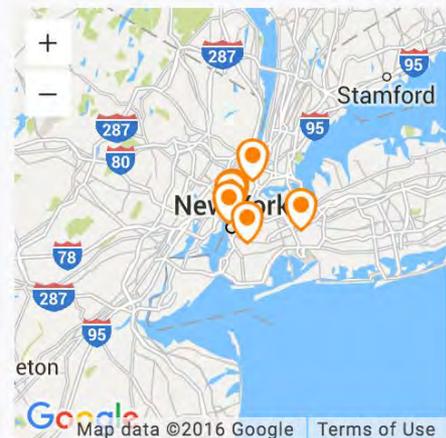
\$44

#filmmedia #festival



SAT, APR 23 5:00 PM

## Tribeca Talks: What We Talk About When We Talk About the Bomb




CATEGORY ▾

EVENT TYPE ▾

DATE ▾

PRICE ▾



https://www.eventbrite

**FETCH & VALIDATE**

CANCEL

Shortlink

Results - Filter by use case ▾

```

1 <!DOCTYPE html>
2 <!--[if lt IE 7 ]> <html class="ie ie6 "
  xmlns="http://www.w3.org/1999/xhtml"
  xmlns:fb="http://ogp.me/ns/fb#" lang="en-us"> <!
  [endif]-->
3 <!--[if IE 7 ]> <html class="ie ie7 "
  xmlns="http://www.w3.org/1999/xhtml"
  xmlns:fb="http://ogp.me/ns/fb#" lang="en-us"> <!
  [endif]-->
4 <!--[if IE 8 ]> <html class="ie ie8 "
  xmlns="http://www.w3.org/1999/xhtml"
  xmlns:fb="http://ogp.me/ns/fb#" lang="en-us"> <!
  [endif]-->
5 <!--[if IE 9 ]> <html class="ie ie9 "
  xmlns="http://www.w3.org/1999/xhtml"
  xmlns:fb="http://ogp.me/ns/fb#" lang="en-us"> <!
  [endif]-->
6 <!--[if (gt IE 9)|!(IE)]><!--> <html class=""
  xmlns="http://www.w3.org/1999/xhtml"
  xmlns:fb="http://ogp.me/ns/fb#" lang="en-us">

```

<b>startDate:</b>	2016-05-02T18:30:00-04:00
<b>name:</b>	Theater Talks: Turn Me Loose
<b>location [Place]:</b>	
<b>name:</b>	Schomburg Center for Research in Black Culture
<b>geo [GeoCoordinates]:</b>	
<b>latitude:</b>	40.81461619999999
<b>longitude:</b>	-73.94090140000003
<b>address [PostalAddress]:</b>	
<b>addressRegion:</b>	NY
<b>addressLocality:</b>	New York
<b>streetAddress:</b>	515 Malcolm X Boulevard
<b>postalCode:</b>	10037
<b>addressCountry [Country]:</b>	
<b>name:</b>	US
<b>organizer [Organization]:</b>	
<b>name:</b>	Schomburg Center for Research in Black Culture
<b>url:</b>	http://www.eventbrite.com/o/schomburg-center-for-research-in-black-c

# ■ Many challenges remain

- Only content of interest to search engines
  - Centralized small schema
- 
- Only big players consume the data
  - Crawl/index is too big a barrier

## ■ Analogy with Web: HTML : ?

- HTML provided small set of standard terms ('div', 'table', 'body', etc.)
- All documents that stuck to these were understood by all browsers
- What is the equivalent here?

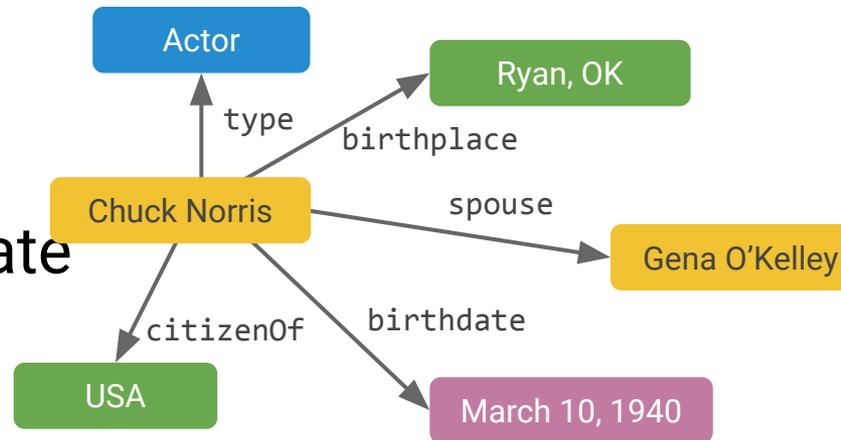
Billions of entities ...

# Game of the name

~1000s of terms like Actor, birthdate

~10s for most sites

Common across sites



~1b-100b terms like Chuck Norris and Ryan, Oklahoma

Cannot expect agreement on these

Need something much more sophisticated than HTML

# Web Analogy: Search engines

- Search engines made web usable
- Need something similar here
  - Collect data from different publishers (Crawl)
  - Aggregate it (Index)
  - Serve (Ranking)
- Consumer here is a program, not human!

# Challenges

- Crawling: hyperlinks help web crawl.
  - What is the analog here?
  - Overlay of web pages that link to datasets?

# ■ Challenges: Building the index

Analog of words : entities

Building the index ---> large scale entity recon

# ■ Challenges: Ranking

Single answer vs ranked set of possible answers

Ranking could be based on authoritativeness of source

# Representation

Simple graph representation is easy to understand

Can't do a lot of things, e.g., time

How rich should the KR lang be?

N-ary rels, negation, quantifiers, ...

Do we have to agree?

# Representation

Hybrid representations

- structured + unstructured
- embeddings

# Concluding

- Many interesting research problems
- Good news:
  - Lot of data already being published
  - First generation of apps already there
  - Don't have to solve all these problems before useful things can be done



Questions?

# Approach is important

Deep pool of AI research to draw from

Attempts to create a web of structured data have fallen short

'Design philosophy' (meta-architecture) important

Gradual investment/learning curves

Permissionless innovation

# Deep pool of research

Representation formalisms from different fields

Mathematical logic

Cognitive Science

Database systems

Statistics, Probabilistic logics

How to represent X: time, actions, defaults, beliefs, ...

# Long history of systems

Advice taker, General problem solver,

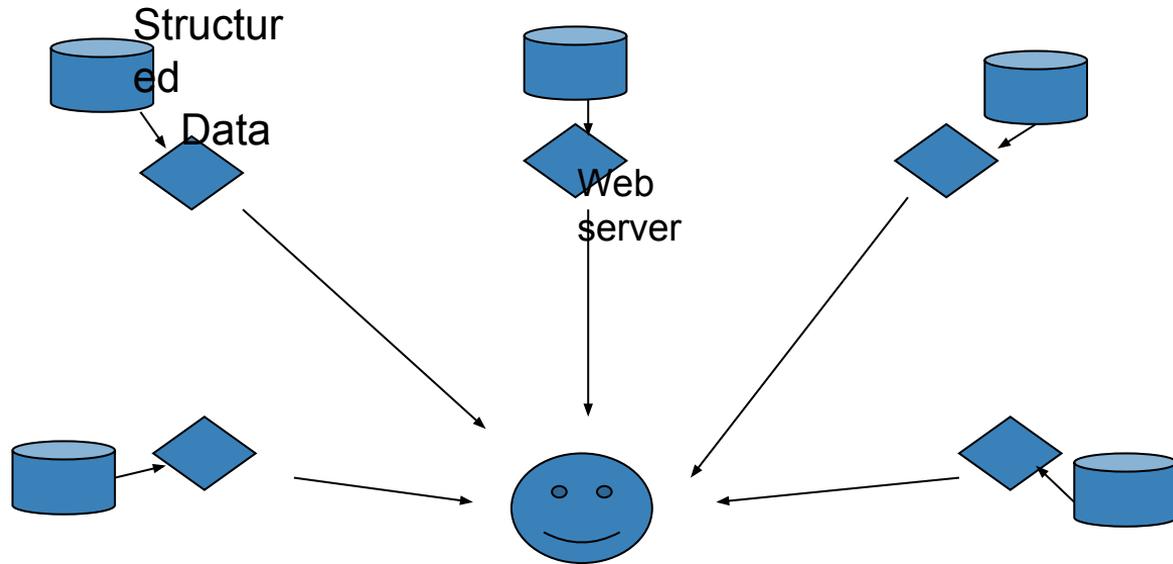
Expert systems, Soar

Cyc

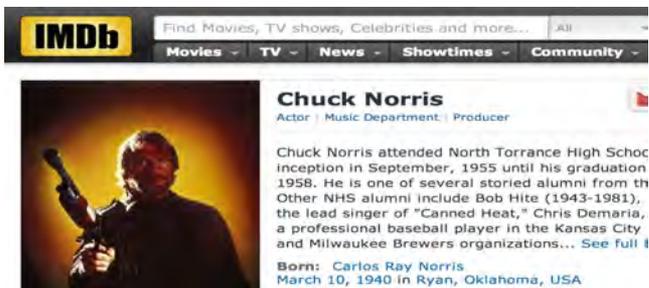
Structured data on the web

# Structured data & Web

Making structured data a first class thing on the web



# The Goal



**IMDb** Find Movies, TV shows, Celebrities and more... All

**Chuck Norris**  
Actor | Music Department | Producer

Chuck Norris attended North Torrance High School in September, 1955 until his graduation in 1958. He is one of several storied alumni from the school. Other notable alumni include Bob Hite (1943-1981), the lead singer of "Canned Heat," Chris Demaria, a professional baseball player in the Kansas City and Milwaukee Brewers organizations... See full bio

**Born:** Carlos Ray Norris  
March 10, 1940 in Ryan, Oklahoma, USA



Celebrities > Chuck Norris

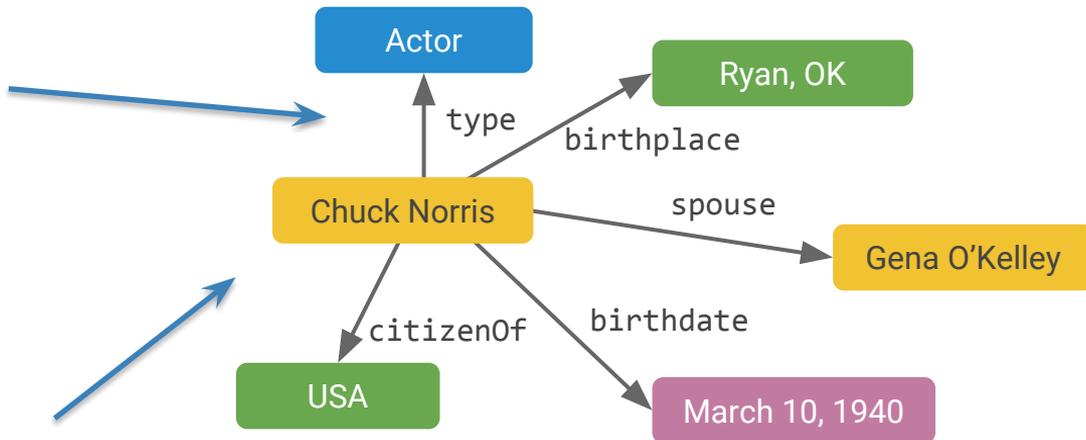
**Chuck Norris**

Highest Rated: 100% Return of the Dragon (The Way of the Dragon) (1973)  
Lowest Rated: 0% Firewalker (1986)

Birthdate: Mar 10, 1940  
Birthplace: Ryan, Oklahoma, USA

Bio: American action star Chuck Norris first learned martial arts while serving in the U.S. Army. From 1968 through 1974, he held the world's middleweight karate championship. In 1975, he made his film debut in The Wrecking Crew (1968) and his TV bow on The Streets of San Francisco. Thanks...

[Full Chuck Norris Bio](#)



Graph Data Model  
Common Vocabulary

# Timeline of efforts

Many attempts: MCF, RDF, OWL, Microformats, OGP, Linked Data,  
...

Some successful, RSS, Vcard, but narrow in scope

Circa 2008, we were beginning to see some adoption, but straight forward copying of web architecture (let a million schemas bloom) was leading to chaos



# Schema.org

Work started in August 2010. Google, Microsoft, Yahoo ...  
Now also Apple, W3C ...

Provides core vocabulary for people, places, events, offers, actions, ... Understood by the search engines

Search (structured data in search) was driving application

# New York Talk Events

RELEVANCE

DATE



MON, MAY 2 6:30 PM

## Theater Talks: Turn Me Loose

Schomburg Center for Research in Black Culture

FREE

#filmmedia #seminar



FRI, APR 22 8:00 PM

## Tribeca Talks After the Movie - Special Correspondents

John Zuccotti Theater @ BMCC Tribeca Performing Arts Center)

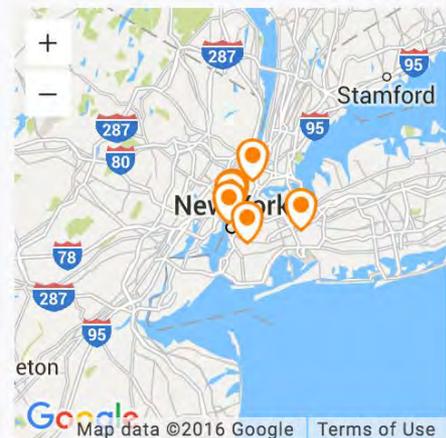
\$44

#filmmedia #festival



SAT, APR 23 5:00 PM

## Tribeca Talks: What We Talk About When We Talk About the Bomb




CATEGORY ▾

EVENT TYPE ▾

DATE ▾

PRICE ▾



https://www.eventbrite

**FETCH & VALIDATE**

CANCEL

Shortlink

Results - Filter by use case ▾

```

1 <!DOCTYPE html>
2 <!--[if lt IE 7 ]> <html class="ie ie6 "
  xmlns="http://www.w3.org/1999/xhtml"
  xmlns:fb="http://ogp.me/ns/fb#" lang="en-us"> <!
  [endif]-->
3 <!--[if IE 7 ]> <html class="ie ie7 "
  xmlns="http://www.w3.org/1999/xhtml"
  xmlns:fb="http://ogp.me/ns/fb#" lang="en-us"> <!
  [endif]-->
4 <!--[if IE 8 ]> <html class="ie ie8 "
  xmlns="http://www.w3.org/1999/xhtml"
  xmlns:fb="http://ogp.me/ns/fb#" lang="en-us"> <!
  [endif]-->
5 <!--[if IE 9 ]> <html class="ie ie9 "
  xmlns="http://www.w3.org/1999/xhtml"
  xmlns:fb="http://ogp.me/ns/fb#" lang="en-us"> <!
  [endif]-->
6 <!--[if (gt IE 9)|!(IE)]><!--> <html class=""
  xmlns="http://www.w3.org/1999/xhtml"
  xmlns:fb="http://ogp.me/ns/fb#" lang="en-us">

```

<b>startDate:</b>	2016-05-02T18:30:00-04:00
<b>name:</b>	Theater Talks: Turn Me Loose
<b>location [Place]:</b>	
<b>name:</b>	Schomburg Center for Research in Black Culture
<b>geo [GeoCoordinates]:</b>	
<b>latitude:</b>	40.81461619999999
<b>longitude:</b>	-73.94090140000003
<b>address [PostalAddress]:</b>	
<b>addressRegion:</b>	NY
<b>addressLocality:</b>	New York
<b>streetAddress:</b>	515 Malcolm X Boulevard
<b>postalCode:</b>	10037
<b>addressCountry [Country]:</b>	
<b>name:</b>	US
<b>organizer [Organization]:</b>	
<b>name:</b>	Schomburg Center for Research in Black Culture
<b>url:</b>	http://www.eventbrite.com/o/schomburg-center-for-research-in-black-c

# Schema.org ... the numbers

Approx. 1800 terms (classes + attributes)

In use by ~15 million sites

- Roughly 30% of pages in search index have markup

- ~25 'triples' per page

- 30% growth over last 12 months

~40% of US/EU ecommerce emails (sales confirmation, reservations, etc.) use schema.org markup

# Schema.org: Major sites

News: Nytimes, guardian, bbc,

Movies: imdb, rottentomatoes, movies.com

Jobs / careers: careerjet, monster, indeed, simplyhired

People: linkedin.com, facebook

Products: ebay, alibaba, sears, cafepress, sulit, fotolia

Local: yelp, allmenus, urbanspoon

Events: wherevent, meetup, zillow, eventful

Music: last.fm, soundcloud

....

# Going beyond Schema.org

We want a much richer, wider range of data

Schema development still too centralized

Too few applications using the data

Two architectural issues

- How much representational richness

- Aggregation (or what is the analog of the hyperlink?)

# Representational Richness, inference

We will want to represent a wide range of phenomenon

50 years of research in structured data representation has given us some wide range of expressive languages

Simple triples are by far the most basic and run out of steam with scale

# Going beyond triples

More expressive representations

More aspects of FOPC (n-ary, negations, ...),  
representations for time, Ontologies, defaults, probabilities,  
distributed Representations

Database vs services

Simple data source vs agent / service

Local smart vs global smart

# Incremental Complexity: Optimizing for flexibility

Delicate tradeoff between ease of use and capability

Optimal point varies with adoption curve

HTML in 1994 vs HTML in 2014

Incrementally introduce complexity

Unclear what the right tradeoff point is

Optimizing for flexibility enables rapid, distributed exploration of the design space

# Coping with distributed

Having millions of data providers has some downsides

- Significant variation in quality

- Many interesting services want to work on aggregates

  - e.g., data mining, search

Two research problems to handle aggregates

- Structured data aggregation is harder than text aggregation

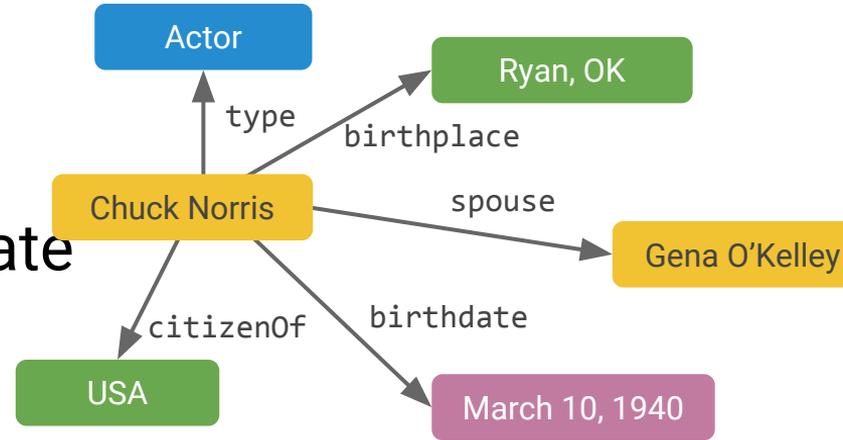
- Creating and working with aggregates: high initial costs

# Game of the name

~1000s of terms like Actor, birthdate

~10s for most sites

Common across sites



~1b-100b terms like Chuck Norris and Ryan, Oklahoma

Cannot expect agreement on these

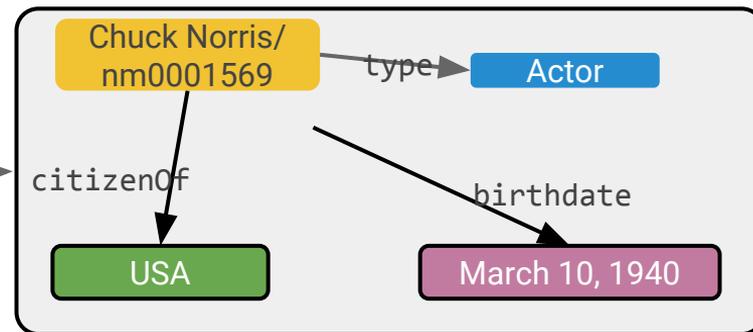
Reference by Description

Consuming applications reconcile entity references

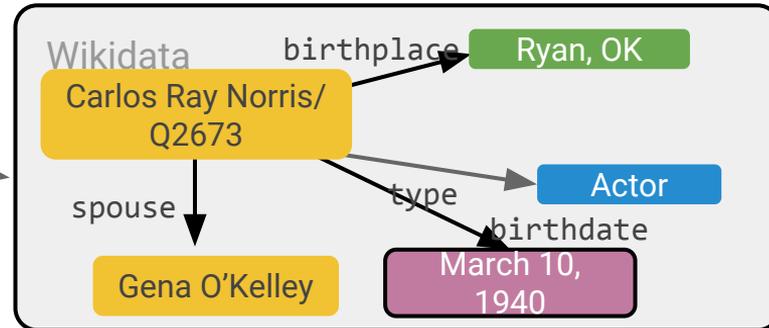
# What Schema.org data looks like

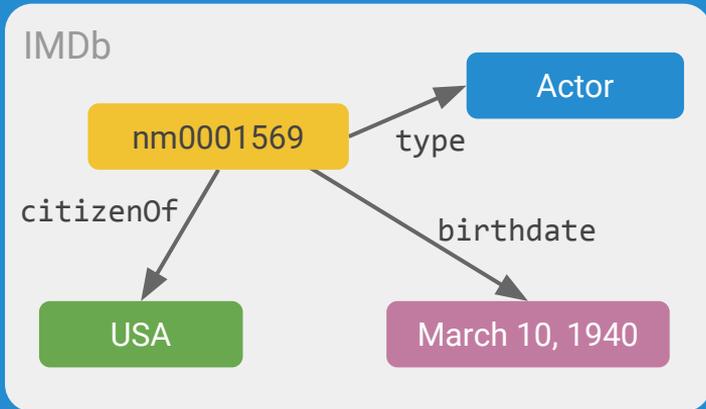


```
<h1 itemprop="name">
  Chuck Norris
</h1> ...
```

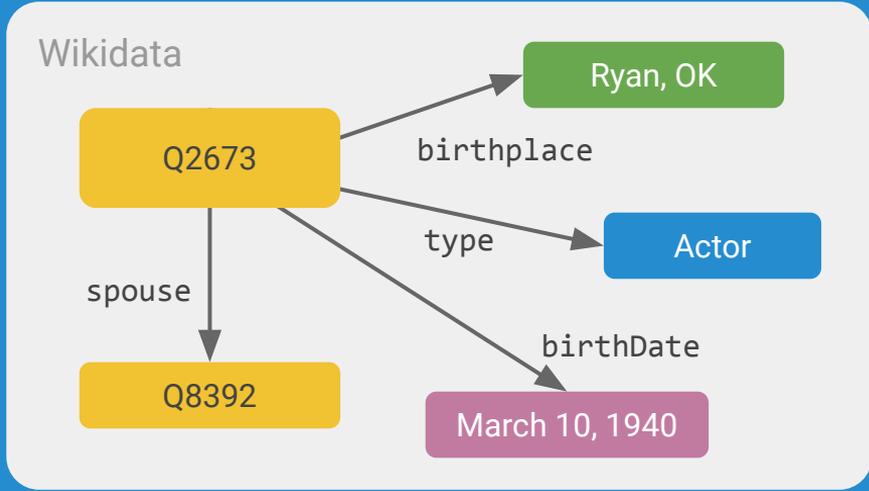


```
<time
  datetime="1940-3-10"
  itemprop="birthDate">
```

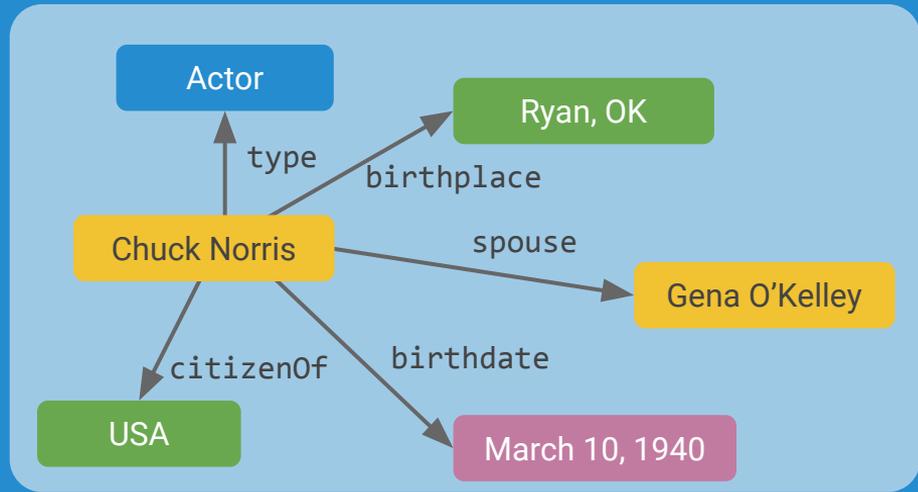




+



=



Stitch

# Aggregate datasets

Research is driven by large, interesting datasets

Datasets open new frontiers

- Genomics
- ImageNet
- Skyscraper: Sloan Digital Sky Survey

Contrast with web search



IMGE



SLOAN DIGITAL SKY SURVEY

# Using Datasets: current model

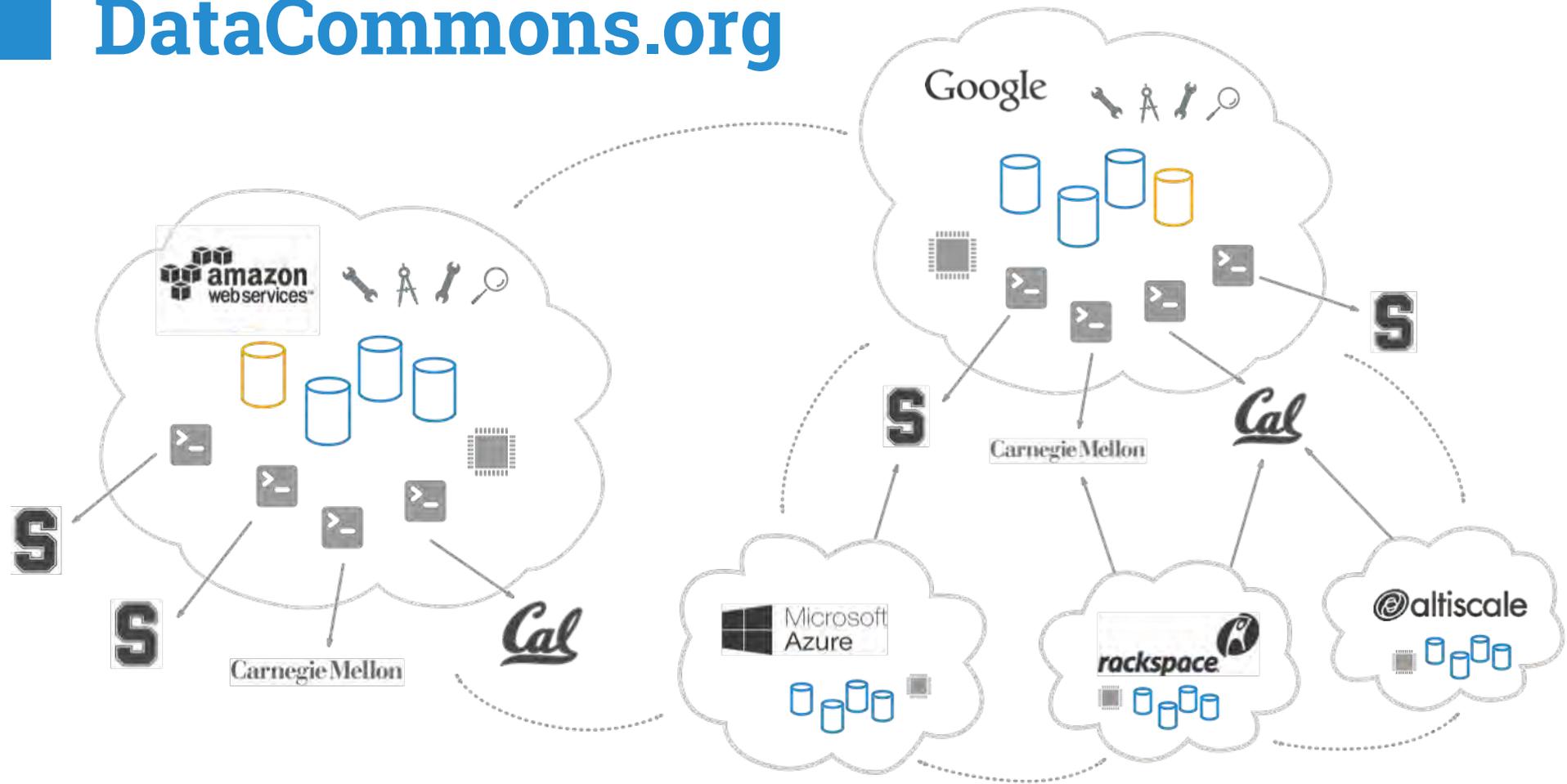
Here is a dataset, download and have fun

High upfront costs: machines, storage

Sparse ecosystem, few tools, ...

Hello world is just too hard!

# DataCommons.org



# Data Commons

Bring the code to the data

Make it much easier to play around with

Hello world → Trying something small should take < 30 min

Will lead to ecosystem of creators and users shared data sets, tools, applications ...



# Concluding

Big opportunity with Knowledge Graphs

We need the kind of breadth and depth the Web has

Anyone should be able to participate: Permissionless innovation

At a low entry cost

The story of Aldus

**Thank you**