Schema.org Update

Guha
Outline of talk

• The context
  – How did we end up where we are with the ‘Semantic Web’

• Schema.org
  – What it is, status of adoption
  – Interesting examples & applications
  – Schema.org principles, how does it work
  – Schemas in the pipeline

• Research problems/opportunities
About 17 years ago, ...

- People started thinking about structured data on the web
  - A few people from Netscape, Microsoft and W3C got together @MIT

- Trying to make sense of a flurry of activity/proposals
  - XML, MCF, CDF, Sitemaps, ...

- There were a number of problems
  - PICS, Meta data, sitemaps, ...

- But one unifying idea
Context: The Web for humans
Goal: Web for Machines & Humans
What does that mean?

Actor

Ryan, Oklahoma

birthplace

Chuck Norris

type

birthdate

March 10th, 1940
How do we get there?

• How does the author give us the graph
  – Data Model: Graph vs tree vs ...
  – Syntax
  – Vocabulary
  – Identifiers for objects

• Why should the author give us the graph?
Going depth first

• Many heated battles
  – Lot of proposals, standards, companies, ...

• Data model
  – Trees vs DLGs vs Vertical specific vs who needs one?

• Syntax
  – XML vs RDF vs json vs ...

• Model theory anyone
  – We need one vs who cares vs what’s that?
Timeline of ‘standards’

• ‘96: Meta Content Framework (MCF) (Apple)
• ’97: MCF using XML (Netscape) → RDF, CDF
• ’99 ––: RDF, RDFS
• ’01 ––: DAML, OWL, OWL EL, OWL QL, OWL RL
• ’03: Microformats
• And many many many more ... SPARQL, Turtle, N3, GRDDL, R2RML, FOAF, SIOC, SKOS, ...

• Lots of bells & whistles: model theory, inference, type systems, ...

But something was missing ...

• Fewer than 1000 sites were using these standards

• Something was clearly missing and it wasn’t more language features

• We had forgotten the ‘Why’ part of the problem

• The RSS story
’07 - :Rise of the consumers

• Yahoo! Search Monkey, Google Rich Snippets, Facebook Open Graph

• Offer webmasters a simple value proposition

• Search engines to webmasters:
  – You give us data ... we make your results nicer

• Usage begins to take off
  – 1000x increase in markup’ed up pages in 3 years
Yahoo Search Monkey

- Give websites control over snippet presentation
- Moderate adoption
  - Targeted at high end developers
  - Too many choices
Google Rich Snippets: Reviews
Google Rich Snippets: Events
Google Rich Snippets: Recipe View

Coconut Cake
324 reviews - 1 hr 55 min
Food Network invites you to try this Coconut Cake recipe from Ina Garten.
Ingredients: butter, sugar, eggs, vanilla extract, almonds, flour, baking ...
www.foodnetwork.com \ Recipes \ Baking - Cached - Similar

Sour Cream Coffee Cake
246 reviews - 1 hr 40 min
Food Network invites you to try this Sour Cream Coffee Cake recipe from Ina Garten.
Ingredients: butter, sugar, eggs, vanilla extract, sour cream, flour, baking ...
www.foodnetwork.com \ Recipes \ Baking - Cached - Similar

Lemon Yogurt Cake
433 reviews - 1 hr 20 min
Food Network invites you to try this Lemon Yogurt Cake recipe from Ina Garten.
Ingredients: flour, baking powder, yoghurt, sugar, eggs, lemon, vanilla extract ...
www.foodnetwork.com \ Recipes \ Baking - Cached - Similar

Beatty's Chocolate Cake
1103 reviews - 1 hr 35 min
Food Network invites you to try this Beatty's Chocolate Cake recipe from Ina Garten.
Ingredients: butter, flour, sugar, cocoa, baking soda, baking powder, milk ...
www.foodnetwork.com \ Recipes \ Baking - Cached - Similar

Lemon Cake
223 reviews - 1 hr 30 min
Food Network invites you to try this Lemon Cake recipe from Ina Garten.
Ingredients: butter, sugar, eggs, lemon, flour, baking powder, baking soda ...
www.foodnetwork.com \ Recipes \ Baking - Cached - Similar

Flag Cake
132 reviews - 1 hr 15 min
Food Network invites you to try this Flag Cake recipe from Ina Garten.
Google Rich Snippets

- Multi-syntax
- Adhoc vocabulary for each vertical
- Very clear carrot
- Lots of experimentation on UI
- Moderately successful: 10ks of sites
- Scaling issues with vocabulary
Situation in 2010

• Too many choices/decisions for webmasters
  – Divergence in vocabularies
    • Too much fragmentation
    • N versions of person, address, ...

• A lot of bad/wrong markup
  – ~25% for micro-formats, ~40% with RDFA
  – Some spam, mostly unintended mistakes

• Absolute adoption numbers still rather low
  – Less than 100k sites
Schema.org

• Work started in August 2010
  – Google, Yahoo!, Microsoft & then Yandex (Baidu, sort of)

• Goals:
  – One vocabulary understood by all the search engines
  – Make it very easy for the webmaster

• It is A vocabulary. Not The vocabulary.
  – Webmasters can use it together other vocabs
  – We might not understand the other vocabs. Others might
Schema.org: Major sites

• News: Nytimes, guardian.com, bbc.co.uk,
• Movies: imdb, rottentomatoes, movies.com
• Jobs / careers: careerjet.com, monster.com, indeed.com
• People: linkedin.com,
• Products: ebay.com, alibaba.com, sears.com, cafepress.com, sulit.com, fotolia.com
• Videos: youtube, dailymotion, frequency.com, vinebox.com
• Medical: cvs.com, drugs.com
• Local: yelp.com, allmenus.com, urbanspoon.com
• Events: whereevent.com, meetup.com, zillow.com, eventful
• Music: last.fm, myspace.com, soundcloud.com
Schema.org: categories

• Most used categories by occurrence

• Most used categories by domains
Schema.org: properties

• Top properties by occurrence
  – name, url, image, description, offers, author, price, thumbnailUrl, datePublished, addressLocality, address, itemOffered, duration, streetAddress, isFamilyFriendly, priceCurrency, playerType, paid, regionsAllowed, postalCode, hiringOrganization, jobLocation,

• Top properties by domain
  – Name, description, url, image, contentURL, address, author, telephone, price, postalCode, offers, ratingValue, priceCurrency, datePublished, addressRegion, availability, email, bestRating, creator, review, location, startDate
Applications

• Applications drive adoption

• First generation of applications
  – Rich presentation of search results

• Many new applications are coming up
  – On search page and beyond
Newer Applications: Knowledge Graph
Newer Applications: Knowledge Graph
Non web search Applications

• Searching for Veteran friendly jobs
Non search applications: Google Now

![Pinterest Homepage](image)

#### Easy Apple Crisp

<table>
<thead>
<tr>
<th>Produce</th>
<th>Cooking &amp; Baking</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 Apples</td>
<td>1/2 cup All-purpose flour</td>
</tr>
<tr>
<td>1/2 Oats, old-fashioned</td>
<td>3/4 cup Brown sugar, packed</td>
</tr>
<tr>
<td></td>
<td>3/4 tsp Cinnamon, ground</td>
</tr>
<tr>
<td></td>
<td>1/4 tsp Cloves, ground</td>
</tr>
<tr>
<td></td>
<td>1/2 tsp Nutmeg, ground</td>
</tr>
</tbody>
</table>

**Dairy**

- 1/3 cup Butter

[Make this on Looks Like Homemade](link)
Non search Applications

• Open Table website ➔ confirmation email ➔ Android Reminder
Schema.org principles: Simplicity

- Simple things should be simple
  - For webmasters, not necessarily for consumers of markup
  - Webmasters shouldn’t have to deal with N namespaces

- Complex things should be possible
  - Advanced webmasters should be able to mix and match vocabularies

- Syntax
  - Microdata, usability studies
  - RDFa, json-ld, ...
Schema.org principles: Simplicity

• Can’t expect webmasters to understand Knowledge Representation, Semantic Web Query Languages, etc.

• It has to fit in with existing workflows

• Avoid KR system driven artifacts
  – domainIncludes/rangeIncludes
  – No classes like ‘Agent’
  – Categories and attributes should be concrete
Schema.org principles: Simplicity

• Copy and edit as the default mode for authors
  – It is not a linear spec, but a tree of examples

• Vocabularies
  – Authors only need to have local view
  – But schema.org tries to have a single global coherent vocabulary
Schema.org principles: Incremental

• Started simple
  – ~ 100 categories at launch

• Applies to every area
  – Add complexity after adoption
  – now ~1200 vocab items
  – Go back and fill in the blanks

• Move fast, accept mistakes, iterate fast
Schema.org Principles: URIs

- ~1000s of terms like Actor, birthdate
  - ~10s for most sites
  - Common across sites

- ~10ks of terms like USA
  - External enumerations

- ~1b-100b terms like Chuck Norris and Ryan, Oklahoma
  - Cannot expect agreement on these
  - Reference by description
  - Consumers can reconcile entity references
Schema.org Principles: Collaborations

• Most discussions on public W3C lists

• Work closely with interest communities

• Work with others to incorporate their vocabularies
  – We give them attribution on schema.org
  – Webmasters should not have to worry about where each piece of the vocabulary came from
  – Webmasters can mix and match vocabs
Schema.org Principles: Collaborations

- IPTC /NYTimes / Getty with rNews
- Martin Hepp with Good Relations
- US Veterans, Whitehouse, Indeed.com with Job Posting
- Creative Commons with LRMI
- NIH National Library of Medicine for Medical vocab.
- Bibextend, Highwire Press for Bibliographic vocabulary
- Benetech for Accessibility
- BBC, European Broadcasting Union for TV & Radio schema
- Stackexchange, SKOS group for message board
- Lots and lots and lots of individuals
Schema.org Principles: Partners

• Partner with Authoring platforms
  – Drupal, Wordpress, Blogger, YouTube

• Drupal 8
  – Schema.org markup for many types
    • News articles, comments, users, events, ...
  – More schema.org types can be created by site author
  – Markup in HTML5 & RDFa Lite
  – Come out early 2014
Recent/Upcoming Vocabularies

• Actions, Fleshing out Events
• Commerce: Orders, Reservations, ...
• Communication: Fleshing out TV, Radio, Email, Q&A, ...
• Media: Scholarly works, Comics, Serials
• Sports
• and many many more ...
Big initiatives underway

• Representing time
  – Lot of triples with associated time interval

• Tabular / CSV data
  – Census data, Scientific data, etc.
  – Need mechanisms for external specification of the meaning of these tables
Research ideas

• There are a large number of projects (e.g., Nell@cmu) that are trying to extract triples from the web

• Schema.org markup == Very large training set
Research Idea: Stich

• Billions of triples sharded across millions of sites
• Lots of common entities, but no cross pointers

• Need to put together the graph
  – Like solving the puzzle
An Actor named Chuck Norris
citizenOf USA
birthdate March 10th 1940

A city named Ryan
in the state OK

An Actor named Chuck Norris
birthplace Ryan, Oklahoma
birthdate March 10th 1940

A Person named Geena O’Kelley
spouse Chuck Norris
birthdate March 10th 1940

= An Actor named Chuck Norris
citizenOf USA
birthdate March 10th 1940

= A Person named Geena O’Kelley
spouse Chuck Norris
citizenOf USA
birthdate March 10th 1940
Lessons from schema.org

• Schema.org has succeeded beyond anyone’s imagination

• Make sure you have your carrot!
  – Carrots work much better than sticks!

• Find the right initial level of generality

• Start simple and iterate fast
This program was brought to you by the generous support from folks at Microsoft, Yandex, Yahoo!, Google and over five million webmasters
Questions?