Light at the end of the tunnel

R.V. Guha
Google
Outline of talk

• How did we end up here ... a personal perspective
  – The tortuous path through standards and products

• Schema.org
  – Why schema.org, principles, how does it work
  – Status : schemas, adoption, partners, applications

• Reports from Google, Bing, Yahoo! & Yandex

• Looking forward: Schemas in the pipeline, Research problems
One day, 16 years ago, ...

- Ralph Swick, Ora Lasilla, Tim Bray, Eric Miller and myself

- Trying to make sense of a flurry of activity
  - 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?

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

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

Actor

Ryan, Oklahoma

type

birthplace

Chuck Norris

birthdate

March 10th 1940
How do we get there?

• How does the author give us the graph
  – Data Model
  – 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, inferencing, type systems, ...
But something was missing ...

- Fewer than 10k 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
Yandex Enhanced Snippets & Services

- Snippets for Web Search
- Data for Vertical Search Services
- Various syntax and vocabularies
- Yandex specific vocabs for unexamined use cases
Google Rich Snippets: Reviews

Xanh Restaurant - Mountain View, CA

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."

XANH Restaurant - Mountain View, CA

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..."

Xanh - Mountain View | Urbanspoon

12 reviews - Price range: $3-$5 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...
Google Rich Snippets: People
Google Rich Snippets: Events

Search results for Britney Spears concert include:

- Britney Spears Concert - Cheap Britney San Jose Tickets
- Britney Spears Tickets - Britney Spears 2011 concert tour
- Britney Spears tickets, concerts and tour dates, Official...
- Britney Spears Tickets, Britney Spears Concert - Britney Tickets...
- Wet Weather Sends Britney Spears Concert Indoors - News Story...

Events:
- Thu, Jun 16: Power Balance Pavillon, Sacramento, CA
- Sat, Jun 18: HP Pavilion at San Jose, San Jose, CA
- Mon, Jun 20: Staples Center, Los Angeles, CA

Check Our Prices. Save 15% or More.
Ticket liquidator.com is rated ★★★★★ 80 reviews
www.ticketliquidator.com
Google Rich Snippets: Recipes

[Google search results for tiramisu recipe]
Google Rich Snippets: Recipe View

in Garten cakes

About 69,700 results (0.22 seconds)

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
1,103 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
• 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 microformats, ~40% with RDFa
  - Some spam, mostly unintended mistakes

- Absolute adoption numbers still rather low
  - Less than 100k sites
In the meantime ...

• The Web has grown
  – >25 million independent active sites
  – Trillions of web pages
  – ~5 billion pages change every day
  – Spam, malware, ...

• In other words: scaling problems along every dimension
Schema.org

- Work started in August 2010
  - Google, Yahoo!, Microsoft & then Yandex

- 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: report card

- Over 15% of all sites/pages now have schema.org
- Over 5 million sites, over 25 billion entity references

- In other words
  - Same order of magnitude as the web
  - Not just ‘promising new stuff’ anymore
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: wherevent.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
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

- ~10m-100m terms like Chuck Norris and Ryan, Oklahoma
  - Cannot expect agreement on these
  - 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 vocab}s
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
  – Coming out early 2014
Schema.org & W3C

• Web Schemas – vocabulary discussion
  Part of W3C Semantic Web Interest Group

• W3C Community Group(s)
  e.g. http://w3.org/community/schemabibex/

• Wiki, Mercurial, email lists, ...
  Informal collaboration rather than classic standardization
  Ideas for extending schema.org welcomed
  Also room this week Thu/Fri, see blog.schema.org
W3C Data Activity

• Subsumes & expands on Sem Web & eGov

• Success for Sem Web in several communities who publish self-describing, re-usable structured data

• schema.org makes it easier for developers to use common vocabulary

• W3C pleased with the success of schema.org and continues to encourage data formats that support mixing of vocabs
Schema.Org Usage @ MS

Today
• Used primarily to support Bing’s Rich Captions
• Also recommended for Windows 8 App Contracts

Tomorrow
• Extended usage in Rich Captions and Bing Search
• Development support via new Bing Platform Dev Center
• Innovative experiences in other Microsoft products
Bing Rich Captions

- Ensure your Schema.Org annotations represent the primary content on the page -> ensures correct caption data is shown
- Ensure your Schema.Org annotations are appropriate and relevant to the content of the page -> improves your chances of being shown
- Rich Captions do not support RDFa 1.1 or JSON-LD -> stick with RDFa 1.0 or Microdata for now
Relevant Links

Bing Platform Dev Center

Bing Webmaster Tools (including Markup Validator)
http://www.bing.com/toolbox/webmaster/

Site Markup Guidance
http://www.bing.com/webmaster/help/marking-up-your-site-with-structured-data-3a93e731
Yahoo! Search

• Rich results
  – Blanco et al. Enhanced Results for Web Search, SIGIR 2011

Parcel 104 - Santa Clara, CA - Yelp, Inc.
www.yelp.com/biz/parcel-104-santa-clara?start=120
★ ★ ★ ★ ★ - (408) 970-6104 - Categories: American (New), Bars.
2700 Mission College Blvd Santa Clara, CA 95054 (408) 970-6104

• Related entities
  – Blanco et al.
    Entity recommendations in Web Search
  – Wed 15:15 Search track
Yahoo! Media

• Schema.org across the Yahoo! Network
  – Article, Photo and Video markup
  – Q&A and Sports under development

• Personalization
  – Based on past reading behavior, Facebook profile, implicit feedback
  – Concepts from a news taxonomy and entity-graph
Yahoo!7

- Linking structured and unstructured content
  - Character bio, appearances, related characters...
  - Articles, tweets, quotes, images, video...
Check in for your flight online. Aeroflot
aeroflot.ru/cms/online_registration
Online registration opens 24 hours and closes 45 minutes before departure. You can check in for your flight if your travel with Aeroflot...

Passenger Information

Name
Surname
Departure City
Flight SU

Check in
Based on Schema.org Actions

Aeroflot check in Moscow

Check in for your flight online. Aeroflot.
aeroflot.ru/cms/online_registration
Online registration opens 24 hours and 45 minutes before departure. You can check in for your flight if your travel with Aeroflot...
Other Products

• Enhanced Snippets for Web Search

• Data for Vertical Search Services (Image Search, Video Search, Ya.Directory, ...)

• Data for algorithms
  (schema.org/PeopleAudience, ...)

Yandex
Links

- Docs: [http://clck.ru/8sA44](http://clck.ru/8sA44)
- Validator: [http://clck.ru/8ridt](http://clck.ru/8ridt)
- Yandex Islands: [http://beta.yandex.com](http://beta.yandex.com)

tilid@yandex-team.ru
ajax@yandex-team.ru
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
Non search Applications

• Open Table website ➔ confirmation email ➔ Android 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 >
Future application

- Clinical trials
  - 4000+ clinical trials at any time in the US alone
  - Almost all the data ‘thrown away’
  - All that gets published is a textual ‘abstract’
- Over half the trials are redundant
  - Earlier trials have the data
  - Assumptions, etc. cannot be re-examined
  - Longitudinal studies extremely hard, but super
Vocabularies in the pipeline

• Actions (Potential Actions), Events
• Accessibility
• 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
  - Hard to force into the triple model
  - Looking at named graphs

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

• Propositional Attitudes
  – Fictional characters
  – Possible actions

• Mapping Entities across sites
  – Billions of entities
  – But hundreds of billions of entity references
  – Very large scale cross site entity mapping
Schema.org: concluding

• Provide webmasters
  – A good reason for adding structured data markup
  – Allow more advanced authors to add much richer markup from many different vocabularies
  – An easy way to do it

• Many different applications emerging

• Many interesting research problems left

• Light at the end of the tunnel
Questions?
MCF

- Introduced the Directed Labeled Graph (DLG) model
- Used to describe site structure
- Provided simple visualization plugin
- ~3m downloads, few thousand sites
  (not too bad for ‘96)
- Documentation was about a page
- Lead to MCF using XML
  (4 pages including diagrams)
97-99: RDF / RDFS

• MCF (4 pages) got renamed to RDF
• Data model becomes a religion
• Lot of new things ... reification, sequences, chains, model theory, ...
• Final version done only by 2002
• RDF Primer is 100 pages long!
• But ... used by very few websites
‘99: RSS 0.91

• Netscape needs customizable home page ala my.yahoo
• No $s to do deals
• Create format and open it up to everyone
  – 2 year target was 5k feeds
  – 14 years later ➔ 100m+ feeds
2002: OWL

- Tim BL, et. al. ‘Semantic Web’ paper
- Answer to lack of adoption: we need more features
- Lots of Darpa and EU funding
- Very powerful solutions ... looking for problems
- ~10 years later ... < 100 sites use it
‘04-07: Microformats

• Reaction to the complexity of RDF & OWL
• Retrofitted vCard, etc. into HTML → hCard
• Adoption by Google & Yahoo! made it wildly successful
• Evolutionary dead-end: no new vocabulary in over 5 yrs.
Lessons

• The RSS story
• Make sure you have your carrot!
  – Carrots work much better than sticks!
• Find the right initial level of generality
• Start simple and iterate fast
• Optimize for flexibility