Light at the end of the tunnel

R.V.Guha

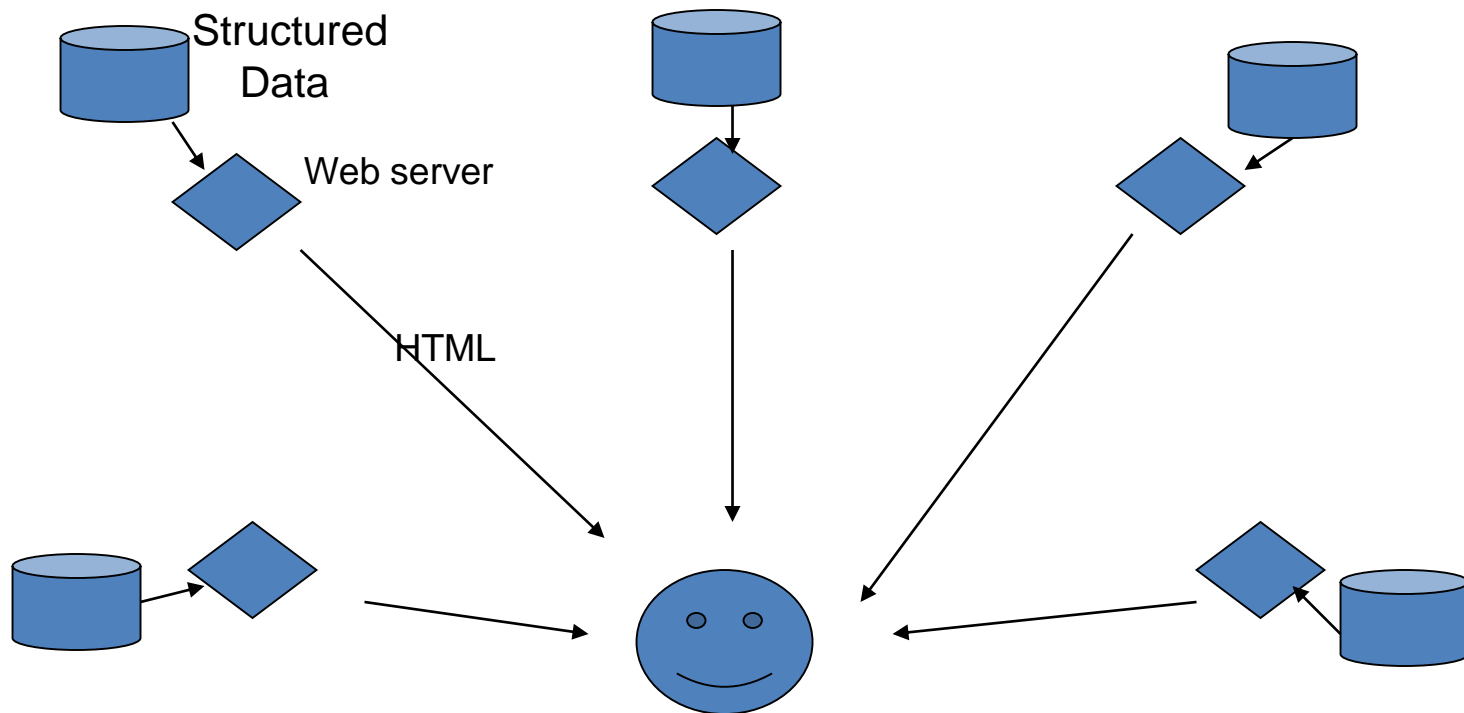Google

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

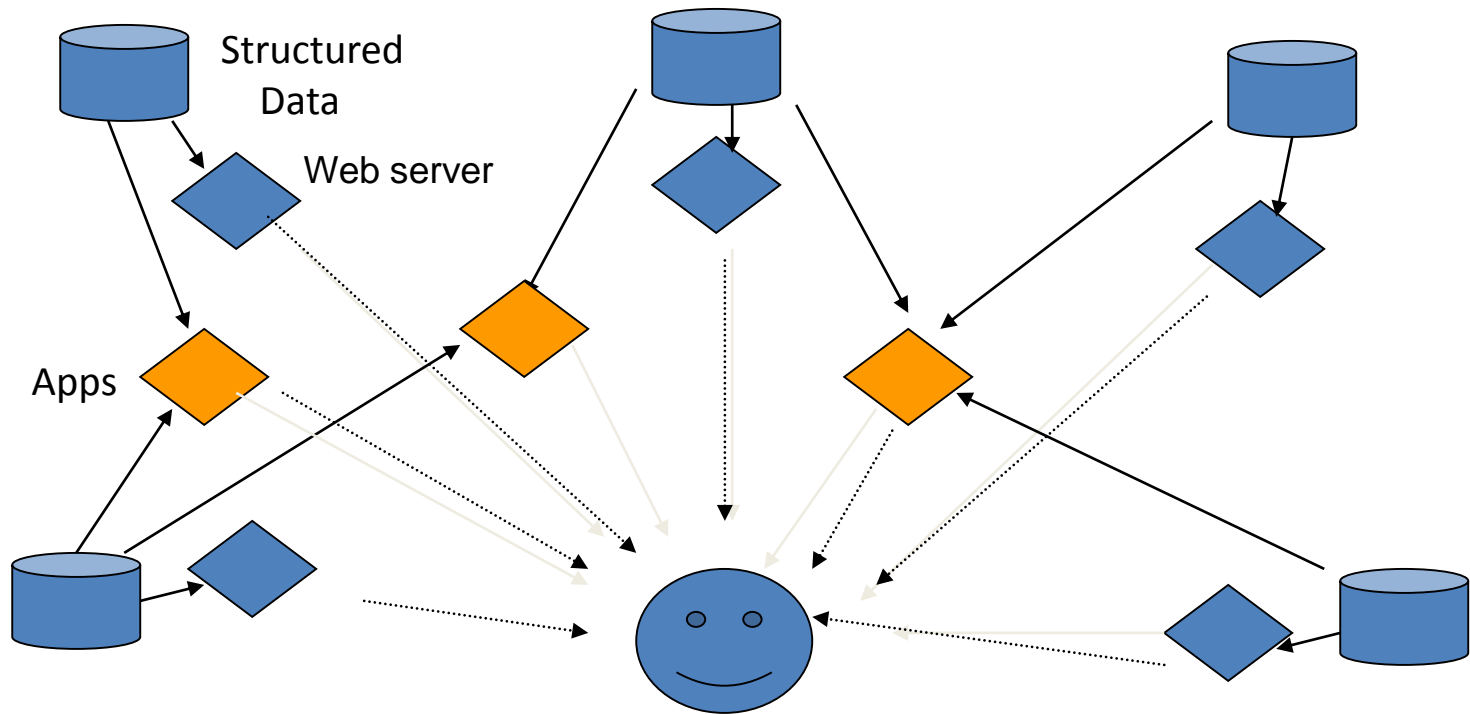- Looking forward: Schemas in the pipeline, Research problems

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
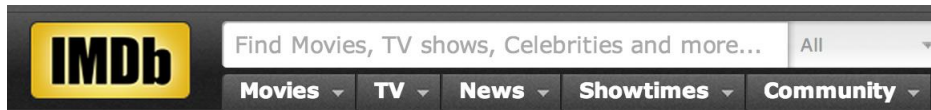 - PICS, Meta data, sitemaps, …

- But one unifying idea

Context: The Web for humans

Structured
Data

Web server

HTML

Goal: Web for Machines & Humans

What does that mean?

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

Tsentr snizheniya vesa i kosmetologii Doktor Bormental
Moscow, Kalashny per., 7
+7 (495) 915-89-89   www.doctorbormental.ru
Cosmetology, Beauty Salons   Psychological Services   Medical surgeries
Directions   Save   Report a problem

Шоколадный торт с молочно-клубничным желе / Кулинарный рецепт
3 яйца, 150 г сахара, 3 ст.л. какао, 100 г муки, 2 ч.л. разрыхлителя, 400 мл клубничного сиропа (компота, сока), 400 мл молока, 40 г желатина, ванильная палочка (или 2 ч...

say7.info › Все кулинарные рецепты › Шоколадный торт с

Google Rich Snippets: Reviews

Google Rich Snippets: People

Google Rich Snippets: Events

Google Rich Snippets: Recipes

Google Rich Snippets: Recipe View

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 <u>A</u> vocabulary. Not <u>The</u> vocabulary.
  - Webmasters can use it together other vocabs
  - <u>We</u> 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

% urls

% urls
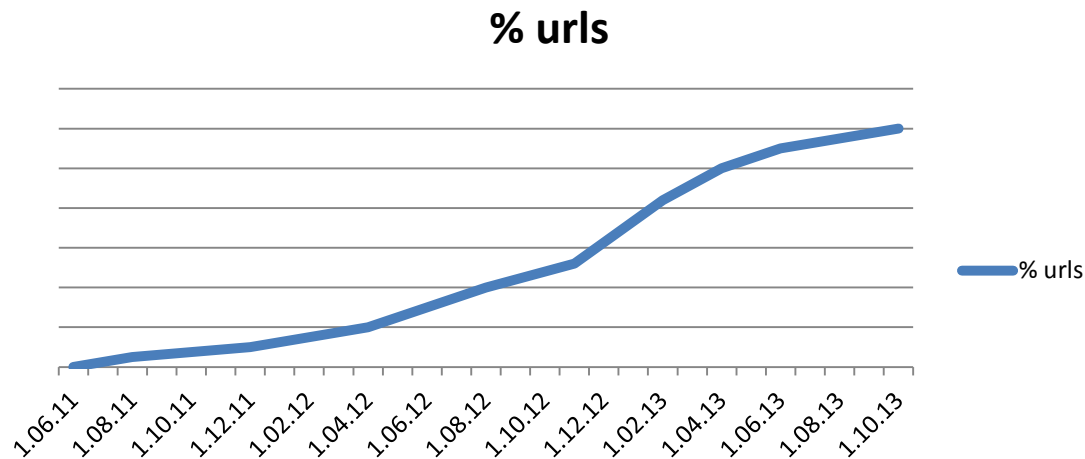
Schema.org: Major sites

- 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
  - Person, Offer, Product, PostalAddress, VideoObject, ImageObject, BlogPosting, WebPage, Article, AggregateRating, LocalBusiness, Place, Organization, MusicRecording, JobPosting, Recipe, Book, Movie, Blog, Photograph, ImageGallery

- Most used categories by domains
  - ImageObject, WebPage, PostalAddress, BlogPosting, Product, Person, Offer, Article, LocalBusiness, Organization, Blog, AggregateRating, Review, VideoObject, Place, Event, Rating, AudioObject, MusicRecording, Store

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, Oklahama
  - Cannot expect agreement on these
  - Consumers can reconcile entity references

Actor

Ryan, Oklahama

type

birthplace

Chuck Norris

birthdate

citizenOf

March 10th 1940

USA
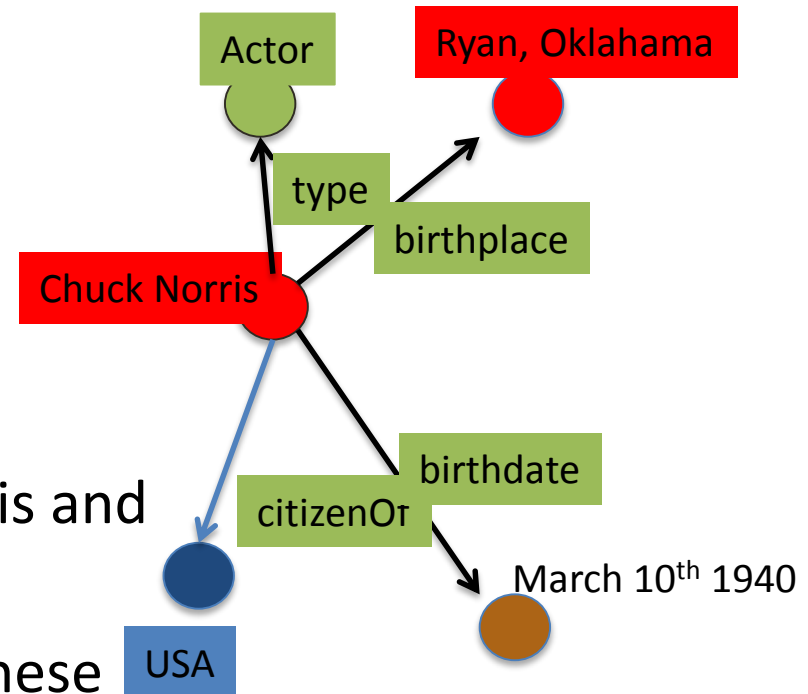
Schema.org Principles: Collaborations

- 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
  - 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

http://www.bing.com/dev/en-us/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
  - Mika & Potter. **Metadata Statistics for a Large Web Corpus,** LDOW 2012

- ## 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
  - *Nicolas Torzec*. **The Y! Knowledge Base: Making Knowledge Reusable at Yahoo!** SemTech 2013

Yahoo!7

- Linking structured and unstructured content
  - Character bio, appearances, related characters…
  - Articles, tweets, quotes, images, video…

Yandex Islands

Yandex Islands

Other Products

- Enhanced Snippets for Web Search

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

- Data for algorithms (schema.org/PeopleAudience, …)

Links

- Docs: http://clck.ru/8sA44

- Validator: http://clck.ru/8ridt

- Yandex Islands: 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
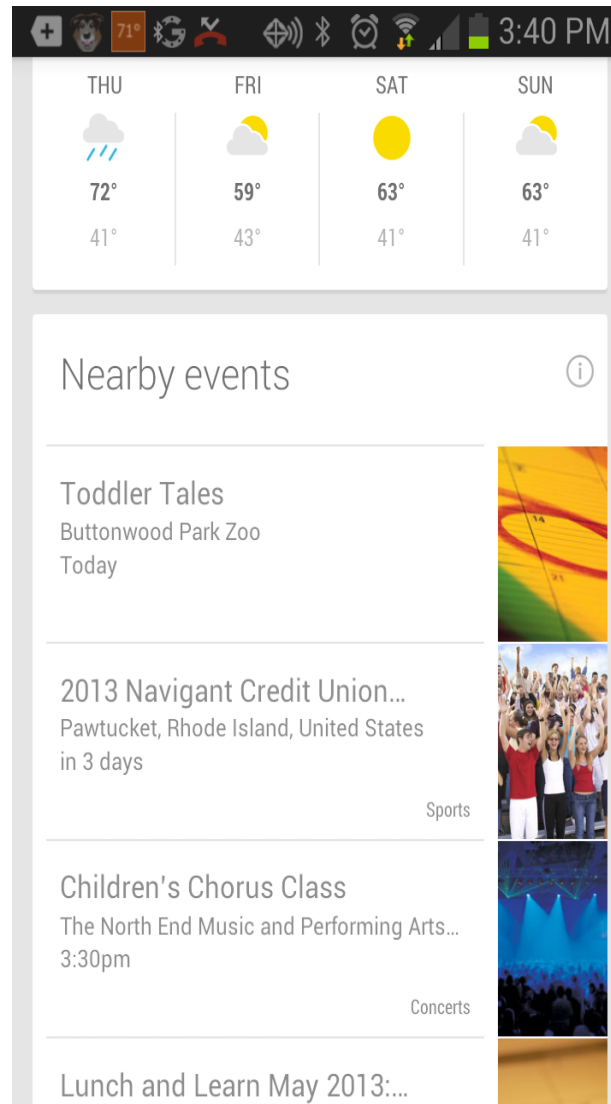
Pinterest: Schema.org for Rich Pins
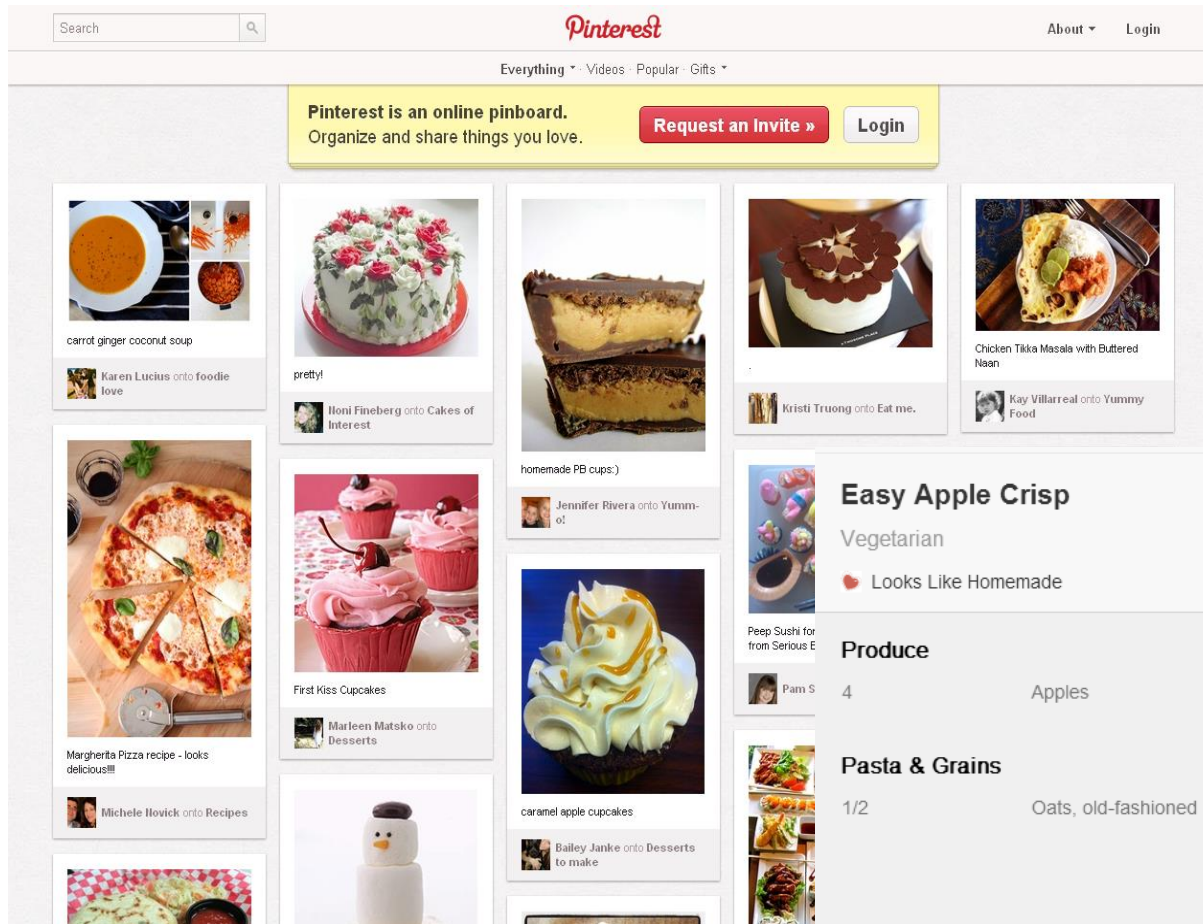
Non search Applications

- 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 >

```
<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>

...
```

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 plug
- ~3m downloads, few thousand si
  (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 <u>initial</u> level of generality
- Start simple and iterate fast
- Optimize for flexibility