Webinar Overview

- Intro: Schema.org, Linked Data & RDF
- A quick look at schema.org
- A longer look at Linked Data and RDF
- Advantages and Challenges of RDF / Linked Data
- Schema.org in detail
- Combining Linked Data and schema.org
- The Future, and what it may hold

Goals

- Show common origins of schema.org and Linked Data
- Illustrate the different design choices available
  - for identifying real-world entities
  - syntaxes for describing things in structured data
  - for integrating multiple vocabularies
  - for incentives to publish (accurate, detailed) data
- usability, expressivity and pluralism
- ...helping guide adoption of schema.org + Linked Data
A quick look.

Introducing schema.org: Search engines come together for a richer web

Today we’re announcing schema.org, a new initiative from Google, Bing and Yahoo! to create and support a common vocabulary for structured data markup on web pages. With schema.org, site owners and developers can learn about structured data and improve how their sites appear in major search engines. The site aims to be a one-stop resource for webmasters looking to add markup to their pages.

Search engines have been working independently to support structured markup for a few years now. We introduced rich snippets to Google search in 2006 to help people find better summaries of reviews and people, and since that time we’ve expanded to new kinds of rich snippets, including recipes and events. We’ve been thrilled to see content creators across the web—via http://schema.org—add markup to their pages, and today we’re able to show rich snippets in search results more than 10 times as often as when we started two years ago.

We want to continue making the open web richer and more useful. We know that it takes time and effort for webmasters to add this markup to their pages, and adding markup is
For example?

```html
<div itemscope itemtype="http://schema.org/VideoObject">
  <h2>Video: <span itemprop="name">My Title</span></h2>
  <meta itemprop="duration" content="T1M33S" />
  <meta itemprop="thumbnailUrl" content="thumbnail.jpg" />
  <meta itemprop="embedUrl" content="http://example.com/videoplayer.swf?video=123" />
  <object type="application/x-shockwave-flash" ...
    <span itemprop="description">Video description</span>
  </object>
</div>
```

Type: http://schema.org/VideoObject
name = My Title
duration = T1M33S
thumbnailurl = thumbnail.jpg
embedurl = http://example.com/videoplayer.swf?video=123
description = Video description

(this is almost all you need to know about RDF, incidentally)
How does it work?

- Normal HTML page content, with additional markup.
- Extra attributes are from W3C HTML5 Microdata: itemscope, itemtype, itemprop, content.
- The HTML5 Microdata syntax was based originally on W3C RDFa, a similar design.
- Both address the principle known on microformats.org as "Don't Repeat Yourself" (DRY), http://microformats.org/wiki/dry.
- Re-use of page markup to also serve as structured data.

Linked Data

Where did it come from...  
...and where is it going?

Linked Data: Prehistory

- Before going deeper, let's go back to the origins of Linked Data publication strategies.
- 'Linked Data' as a slogan and community began with a Tim Berners-Lee note (July 2006) which inspired a Linking Open Data community group and who in turn inspired many others to publish linked, structured data.
- What was new about the 'Linked Data' recipe? How did it improve upon previous RDF practice 1997-2005? How does it compare to the structured data style of schema.org? How did we move from having only a handful of linked RDF files in 2002, through millions to today's talk of billions?

Tim Berners-Lee (2006)

Linked Data

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL, ...).
4. Include links to other URIs so they can discover more things.

http://www.w3.org/DesignIssues/LinkedData
From FOAF to Linked Data

- TimBL's 'Linked Data' note critiqued FOAF's style of linking, since FOAF 2000-2005 did not encourage use of 'http:' identifiers for people and real-world things.

- Instead, FOAF files used a "reference by description" style, describing people in terms of their properties, with "seeAlso" links into a Web of linked RDF documents.

- "This [Friend of a Friend] linking system was very successful, forming a growing social network, and dominating in 2006, the linked data available on the web." (TimBL 2006)

- "However, the system has the snag that it does not give URIs to people, and so basic links to them cannot be made."

- Post-2007 (with supporting changes in the standards) the RDF community adopted this new style, and started publishing RDF with "http:" URIs for real-world things.

Linked RDF: Growth

- In 2002, an early FOAF crawl found RDF about 24 people.

- By 2006 (see Jennifer Golbeck 'Web-based Social Networks A Survey and Future Directions') we saw:
  - LiveJournal.com: 7,500,000 FOAF profiles
  - Ecademy.com: 72,000 FOAF profiles
  - Trust project: 1,700 FOAF profiles
  - ...others, totalling 6 million+ site-based profiles
  - plus smaller number of "in the wild" FOAF files

Linked Data

- The Linking Open Data community took this further, and quickly: publishing structured data from database-backed sites allows very rapid growth.

- LOD data was primarily in RDF/XML (rather than RDFa)

- The big shift: publishing structured data from databases

- LOD data typically assigned http: URIs to real-world things

- LOD structured data was rarely constrained by need to fit human-facing HTML

- LOD structure typically authored by RDF experts using a diverse mix of several RDF vocabularies
The Big Shift

- LOD’s explicit focus on publishing entire datasets boosted adoption massively.
- Assigning Web identifiers to real-world things made data merging and linking much easier than in earlier work.
- The cloud diagram inspired other outside the core RDF community; proof-of-concept conversions evolved into 'officially hosted' participation (e.g. LCSH.info & id.loc.gov)
- Technical publication details remained 'fiddly' (HTTP redirects, debates about URI structure) but an energetic community helped early adopters find their way.
- Buzz around "Web 2.0", APIs and open data fed this growth.
The Basic Idea: **Links** and **claims**

- In Linked RDF (both FOAF and Linked Data):
  - Pages (documents) express claims *about the world*.
  - **Linked documents** describe Linked entities.
  - This idea dates from the earlier days of the Web.
  - Understanding this gives an understanding of the entire design space we are exploring with Linked Data, schema.org and RDF-based languages.
  - *Many technical issues are clearer when you realise there are two kinds of links involved...*

**RDFWeb intro**

June 2000:

The basic idea behind FOAF is simple: the Web is all about making connections between things. FOAF provides some basic machinery to help us tell the Web about the connections between the things that matter to us.

Thousands of people already do this on the Web by describing themselves and their lives on their home page. Using FOAF, you can help machines understand your home page, and through doing so, learn about the relationships that connect people, places and things described on the Web. FOAF uses W3C's RDF technology to integrate information from your home page with that of your friends, and the friends of your friends, and their friends.

FOAF is best explained with an example. Consider a Web of inter-related home pages, each describing things of interest to a group of friends. Each new home page that appears on the Web tells the world something new, providing factoids and gossip that make the Web a mine of disconnected snippets of information. FOAF provides a way to make sense of all this. Here’s an example, a fragment from the mostly-fictional FOAF database. First we list some facts, then describe how the FOAF system makes it possible to explore the Web learning such things.

Dan lives in Zetland Road, Bristol, UK with Libby and Craig. Dan’s email address is danbri@w3.org. Libby’s email address is libby.miller@bris.ac.uk. Craig’s email address is craig@netgates.co.uk. Libby and Craig work for an organisation called “ILRT” whose website is at http://ilrt.org/. Craig works for Netgates, an organisation whose website is at http://www.netgates.co.uk/. Craig’s wife Liz Lives in Bristol with Kathleen, Kathleen and Liz also work at Netgates. Damian lives in London. Martin knows Craig, Damian, Dan and Libby quite well. Martin lives in Bristol and has an email address of m.l.poulter@bristol.ac.uk. (etc...)

This kind of information is the sort of thing typically found on Web home pages. The extract shown here indicates how short, stylised factual sentences can be used to characterise a Web of relationships between people, places, organisations and documents. In real life, this information would be most likely be distributed across various Web pages created by the individuals listed. Very likely, their pages will link directly or indirectly to the home pages of countless other friends-of-friends-of-friends.

**Claims**

- Many pages are about some particular thing, e.g. a person or a movie or book.
- They describe various properties of that thing, often in terms of links to other things; these are **claims** that can be treated as structured data.
- RDFa, Microdata and Microformats explore this in HTML, Linked Data mostly in other formats.
Two kinds of link
Links

- Links between documents, links between things.
- As we compare Linked Data and schema.org, we’ll see different tradeoffs being made.
- Linked Data: an emphasis on richness of description and linkage, over human-readability and webmaster usability.
- Schema.org: simplifies expressivity and weaker linkage, as a tradeoff for greater webmaster/publisher usability.
- These are practical differences of emphasis and focus, rather than conflicts or different architectures.
- Both styles share the RDF data model, and can be blended together in several ways.

But why?

- Let’s now take a few minutes to talk and think about incentives for publishing structured data.
- The incentives for publishing Linked RDF have matured over the years, and this has some effect on our technical choices...
- Who will be publishing, with what tooling, and for what reason?
INCENTIVES
Serious Linked Data

Incentives

- FOAF developer appeal 2000-7: fun demos; networks are intrinsically inspirational; general love of open data, APIs. Publisher appeal: it was easy and futuristic.


- The data itself became a major asset; supporting technology was a means to an end.

Incentives: schema.org

- A very mainstream message relevant to maintainers of millions of Web sites
- To sustain massive non-expert influx, some dramatic simplifications
ETERNAL SEPTEMBER?

"Eternal September (also September that never ended)\cite{1} is a Usenet slang expression, coined by Dave Fischer, for the period beginning September 1993.

It's moot now. September 1993 will go down in net.history as the September that never ended.

"The expression encapsulates the belief that an endless influx of new users (newbies) since that date has continuously degraded standards of discourse and behavior on Usenet and the wider Internet."

Is schema.org the Eternal September of the Semantic Web?

Compare and Contrast

- Classic 'Linked Data' carefully distinguishes identifiers for things, and for pages about those things.
- Schema.org markup tends to leave the 'thing' identifier implicit, for simpler HTML.

Comparison (2.)

- Classic Linked Data: values rich, highly interlinked descriptions that use different independently created vocabularies.
- Schema.org: 300+ classes, 300+ properties in a single schema, addressing many more use cases without mixing vocabularies.
Schema.org

- Search engine collaboration:
  - Google, Bing, Yahoo! & Yandex
- **Simple factual data for better search**
- Launched June 2011, schema.org schema
- 300 classes, 261 properties & growing
- discussions: W3C WebSchemas group

On IMDB:

```html
<div id="content-2-wide" itemscope itemtype="http://schema.org/CreativeWork">
  <div class="txt-block">
    <h4 class="inline">Stars:</h4>
    <a onclick="(new Image()).src='/rg/title-overview/star-1/images/b.gif?link=%2Fnm0010930%2F';" href="/name/nm0010930/" itemprop="actors">Douglas Adams</a>,
    <a onclick="(new Image()).src='/rg/title-overview/star-2/images/b.gif?link=%2Fnm0048982%2F';" href="/name/nm0048982/" itemprop="actors">Tom Baker</a> and
  </div>
  <div class="txt-block">
    Linked Data: see http://www.imdb.com/name/nm0010930/ for schema.org markup describing Douglas Adams as a http://schema.org/Person (jobTitle, birthDate, description, performerIn...).
  </div>
</div>
```

Example: Google Rich Snippets

![Google Rich Snippets](image)

From: http://www.google.com/webmasters/tools/richsnippets
See also Yandex’s http://webmaster.yandex.ru/microtest.xml

What’s in the schema?

- **Classes** (types) e.g. LocalBusiness, Person, Organization, VideoObject, TVSeries...
- **Properties** (attributes) e.g. openingHours, transcript, productionCompany, streetAddress
- That’s all - a dictionary of terms, used for annotating data within normal Web pages
Another example:

```html
<div itemscope itemtype="http://schema.org/Restaurant">
  <span itemprop="name">GreatFood</span>
  <div itemprop="address" itemscope itemtype="http://schema.org/PostalAddress">
    <span itemprop="streetAddress">1901 Lemur Ave</span>, <span itemprop="addressLocality">Sunnyvale</span>, <span itemprop="addressRegion">CA</span>, <span itemprop="postalCode">94086</span>
    <span itemprop="telephone">(408) 714-1489</span>
    <a itemprop="url" href="http://www.dishdash.com">www.greatfood.com</a>
  </div>
  <meta itemprop="openingHours" content="Mo-Th 17:00-21:30">Mon-Thu 5pm-9:30pm <meta itemprop="openingHours" content="Fr-Sa 17:00-22:00">Fri-Sat 5pm-10:00pm
  Categories: <span itemprop="servesCuisine">Middle Eastern</span>, <span itemprop="servesCuisine">Middle Eastern</span>, 
</div>
```

Schema.org scope

- **In-page** structured data for search
- Not asking an unconstrained “so, how do we describe cars?”, but “how can we improve markup on existing pages that describe cars?” (or Comics, SoftwareApps, Sports, ...)
- Simplify publisher/webmaster experience
- Record agreements between search engines
- Central use case: augmented search results
Lots missing! e.g. sports

- Current vocabulary emphasizes 'points of interest' on a map and sporting activities rather than sports content 'as entertainment'
- We also have terms to describe videos, TV shows etc., ...but no sports-specifics yet
- How deep to go? How to integrate with existing vocabulary? How to identify players, teams, kinds of 'football'? Video clips for that 'hand of God' goal?

Everything overlaps

- We added JobPosting; what if the job was sports-related?
- We're adding educational markup; does it help describe sports education, training?
- Is there a sports perspective on the health/medical vocabulary we're working on?
- Can't coordinate everything! Pragmatism...

Practicalities

- Delegation to external sources for enumerations and detail
- e.g. country codes from UN FAO or Wikipedia/DBpedia/Wikidata
- We don’t want to create big enumerations
- all the countries? sports? things that go on maps?
- Decentralised subclassing & property values

* ‘intertwingularity’
Process

• Search partners retain ultimate oversight
• W3C hosts community group, discussion, wiki and proposal tracking
• Web Schemas group - planning monthly telecons at W3C, based around proposals
• Evolving, pragmatic, collaborative

Schema.org

• Shares roots with decentralized Linked Data, including RDF graph data model
• Can be seen as a form of Linked Data
• Page-based markup and mainstream adoption require emphasis on simplicity
• Pragmatic, evolving and co-operative approach

So what's different?

(this is almost all you need to know about RDF, incidentally)
Information Management: A Proposal
by Tim Berners-Lee

Schema.org and Linked Data design landscape

use Web page URLs and omit or conflate with entity identifiers
one big schema that covers many common use cases
HTML-based syntax with minimal new things to learn

syntax-agnostic tools (so tend to ignore usability aspects of each syntax)
expressive mix of schemas that give rich, detailed descriptions
always give URLs for entities & different URLs for docs

easier for consumers
Historically...

- the Linked Data community improved on the early FOAF/RDFWeb linking model by adding explicit URI identifiers for entities.
- Schema.org's need for publisher simplicity softens this guideline, making more work for consumers but more data for everyone.

Practical Convergence

- Using "External enumerations" to add detail to schema.org descriptions - subtypes and property values.
- Using separate Linked Data files to publish rich detail.
- Using Linked Data tooling to consume schema.org data and mix with other RDF.

Common Questions

- Q: Should we stop all our modeling work and just use schema.org since it is understood by the search engines?
- A: No! But consider using schema.org too...

Why model?

- Rich models (e.g. FRBR/RDA, CIDOC, or in lifesciences) have value beyond search. Understand your data and problems!
- You can publish more detailed RDF Linked Data in parallel.
- There's no single 'right' description; if you've an audience, share the data...
Let’s Share What We Know

World Wide Web