# RDF and Semantic Web

## can we reach escape velocity?

---

Jeni Tennison
jeni@jenitennison.com
http://www.jenitennison.com/blog/

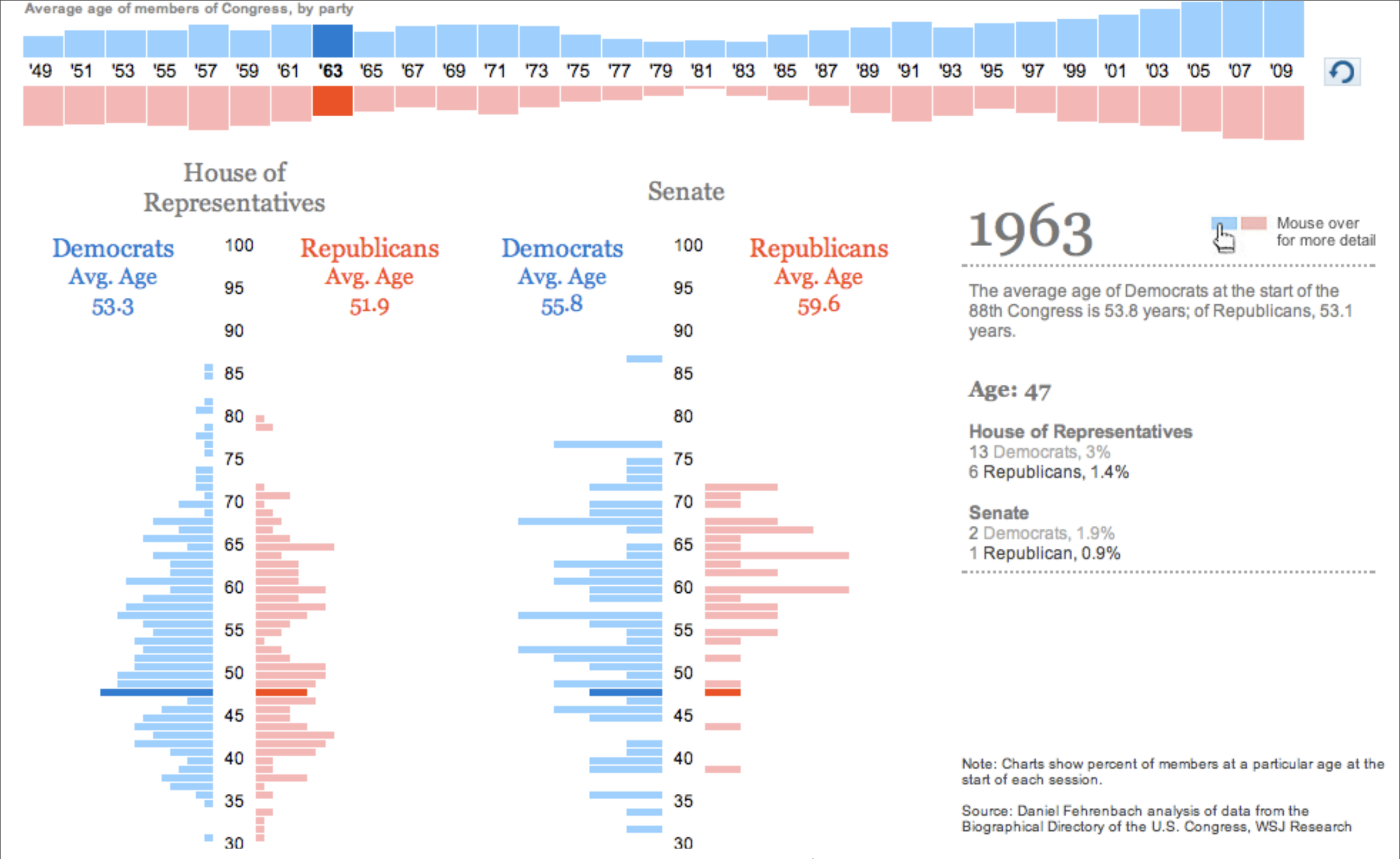
linked data adviser to data.gov.uk
 - not a Semantic Web evangelist!
 - like a lot of people, made the decision 12 years ago that it was all a pipe dream and chose to focus on XML instead

view my role as making sure we're not using linked data just because TimBL says so
 - but equally, giving RDF a fighting chance
 - seeing if it can be appealing and useful within & outside government

given the title for this talk
don't particularly care if RDF & Semantic Web reaches escape velocity
 - do care about making data usable
 - do care about distributed data publication across large, diverse organisations like government
    - like the wider web, but with more committees

three parts
 - describe the potential of the web of data
 - describe where the hurdles are for RDF being used within it
 - explore where W3C should be focusing its efforts

basically taken this as an excuse to have a bit of a rant
 - say some things I wouldn't normally say to people's faces




# Data is hot

## visualisations and APIs

potential of the web of data

world is different from ten years ago when I discounted RDF
open data is exploding
 - APIs on websites
 - publication of open data, particularly in government
 - visualisation and data blogs
 - Strata conference

drive to support third-party reuse
 - apps, widgets
 - visualisation tools
 - taking advantage of funky HTML5 goodness

many visualisations are a lot like pre-web documents
 - rudimentary views with no chance to go deeper

heading towards deeper interactions with data
 - explore, slice, visualise in different ways
 - find out more about the things it refers to
 - data that you can click on as revolutionary as documents you can click on

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Unweighted base – household crime* | *10,905* | *10,059* | *16,310* | *14,900* | *32,720* | *36,395* | *44,973* | *47,610* | *47,027* | *46,765* | *46,252* | *44,610* | |
| **Theft from the person** | **434** | **438** | **680** | **621** | **604** | **690** | **584** | **576** | **574** | **581** | **725** | **525** | **-23** |
| Snatch theft from person | 86 | 79 | 80 | 83 | 74 | 88 | 92 | 71 | 72 | 80 | 103 | 64 | *-20* |
| Stealth theft from person | 348 | 359 | 600 | 538 | 529 | 602 | 492 | 504 | 502 | 501 | 622 | 461 | *-23* |
| **Other theft of personal property** | **1,586** | **1,739** | **2,069** | **1,935** | **1,407** | **1,344** | **1,154** | **1,196** | **1,141** | **987** | **1,096** | **1,036** | ***-50*** |
| **All violence** | **2,074** | **2,556** | **4,176** | **3,593** | **2,728** | **2,714** | **2,320** | **2,349** | **2,471** | **2,200** | **2,114** | **2,087** | ***-50*** |
| Wounding | 508 | 624 | 914 | 804 | 648 | 709 | 577 | 547 | 578 | 477 | 466 | 501 | *-45* |
| Assault with minor injury | 609 | 784 | 1,356 | 1,198 | 709 | 623 | 629 | 572 | 571 | 492 | 533 | 428 | *-68* |
| Assault without injury | 793 | 966 | 1,567 | 1,257 | 1,015 | 1,079 | 860 | 918 | 1,002 | 917 | 844 | 823 | *-48* |
| Robbery | 164 | 182 | 339 | 334 | 356 | 303 | 255 | 311 | 320 | 315 | 272 | 335 | *-1* |
| Violence with injury | 1,194 | 1,441 | 2,408 | 2,184 | 1,497 | 1,441 | 1,300 | 1,227 | 1,270 | 1,063 | 1,116 | 1,065 | *-56* |
| Violence without injury | 881 | 1,115 | 1,768 | 1,409 | 1,231 | 1,273 | 1,020 | 1,121 | 1,201 | 1,137 | 998 | 1,021 | *-42* |
| Domestic violence | 292 | 534[7] | 989 | 814 | 626 | 506 | 401 | 357 | 407 | 343 | 293 | 290 | *-71* |
| Acquaintance | 774 | 1,043[7] | 1,816 | 1,642 | 862 | 949 | 828 | 817 | 845 | 776 | 691 | 679 | *-63* |
| Stranger | 844 | 797[7] | 1,004 | 784 | 883 | 956 | 836 | 863 | 894 | 766 | 852 | 783 | *-22* |
| Mugging *(robbery + snatch theft)* | 250 | 259[7] | 419 | 417 | 430 | 391 | 347 | 382 | 392 | 394 | 375 | 398 | *-5* |
| *Unweighted base – personal crime* | *10,905* | *10,059* | *16,337* | *14,937* | *32,787* | *36,450* | *45,069* | *47,729* | *47,138* | *46,903* | *46,220* | *44,559* | |
| **All acquisitive crime[8]** | **6,418** | **10,009** | **12,148** | **10,587** | **7,642** | **7,394** | **6,129** | **6,047** | **6,040** | **5,540** | **5,977** | **5,427** | ***-55*** |
| Household acquisitive crime | 4,234 | 7,651 | 9,060 | 7,697 | 5,275 | 5,057 | 4,136 | 3,965 | 4,005 | 3,657 | 3,883 | 3,531 | *-61* |
| Personal acquisitive crime | 2,184 | 2,358 | 3,088 | 2,891 | 2,367 | 2,337 | 1,993 | 2,082 | 2,035 | 1,883 | 2,094 | 1,895 | *-39* |
| **ALL HOUSEHOLD CRIME** | **6,947** | **10,410** | **12,426** | **10,562** | **7,879** | **7,592** | **6,645** | **6,632** | **6,923** | **6,282** | **6,583** | **5,939** | ***-52*** |
| **ALL PERSONAL CRIME** | **4,094** | **4,733** | **6,925** | **6,149** | **4,739** | **4,748** | **4,058** | **4,120** | **4,186** | **3,768** | **3,936** | **3,648** | ***-47*** |
| **ALL BCS CRIME[9]** | **11,041** | **15,142** | **19,351** | **16,712** | **12,618** | **12,341** | **10,703** | **10,752** | **11,109** | **10,050** | **10,518** | **9,587** | ***-50*** |
| *Unweighted base – personal crime* | *10,905* | *10,059* | *16,337* | *14,937* | *32,787* | *36,450* | *45,069* | *47,729* | *47,138* | *46,903* | *46,220* | *44,559* | |

1. For an explanation of year-labels see 'Conventions used in figures and tables' at the start of this volume.
2. The numbers are derived by multiplying incidence rates by the population estimates for England and Wales, that is: for household crimes, by 23,525,137 households and for personal crimes, by 44,647,810 adults. F
3. Prior to 2001/02, BCS estimates relate to crimes experienced in a given calendar year. From 2001/02 onwards the estimates relate to crimes experienced in the last 12 months based on interviews in the given finan
4. Estimates of the total number of households in England and Wales in 2004/05, 2005/06, 2006/07, 2007/08 and 2008/09 have been revised. Estimates of the number of household crimes for these years will differ fro
5. BCS estimates from interviews in 2008/09 have been revised based on revised LFS microdata and may vary slightly from previously published estimates. See Section 8 of the User Guide for more information.
6. Percentage changes for crimes such as snatch theft, robbery and domestic violence should be treated with caution because the number of victims interviewed is low (around 200 in 2009/10).
7. The 1991 estimates for domestic, acquaintance and stranger violence and mugging were calculated based on the estimate for all violence. Estimates for these individual categories could not be calculated using the calculating these rates were not collected for that year.
8. It is not possible to calculate whether a change in all acquisitive crime is statistically significant. Changes in both all personal acquisitive crime and all household acquisitive crime in the same direction indicate that t
9. Statistical significance for change in all BCS crime cannot be calculated in the same way as for other BCS figures (a method based on an approximation has been developed). For more information see Section 8 of
10. See Section 5 of the User Guide for more information about the crime types included in this table.
11. Figures for BCS years not presented in this table are included in an extended version of the table, available online at http://www.homeoffice.gov.uk/rds/crimeew0910.html

# Delving deeper

## understanding & finding more

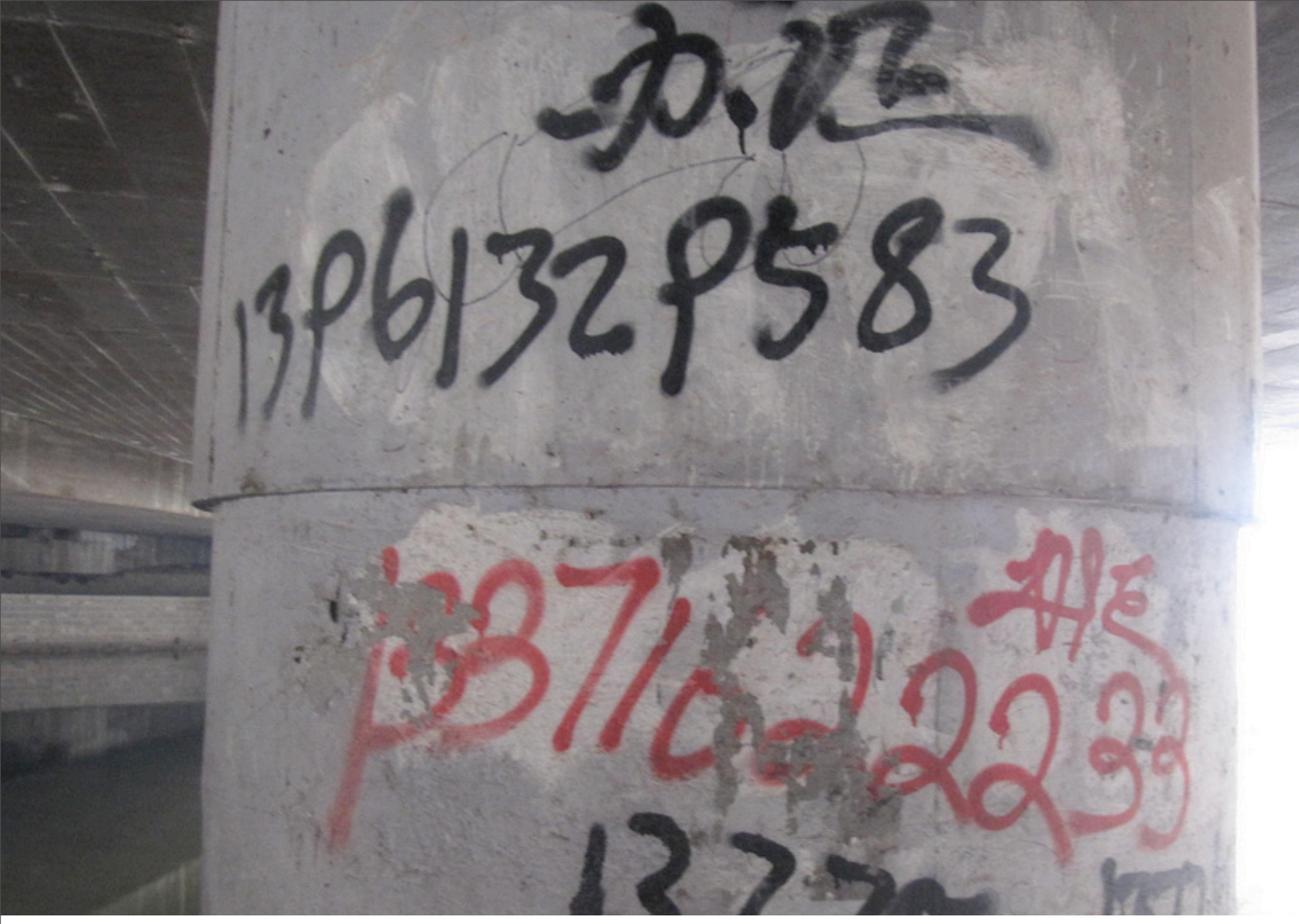
challenges of web of data

first challenge: having data does not mean you understand it
particularly true with government data
example here is UK crime statistics
 - published in Excel because CSV is too inexpressive a medium
what do these rows relate to?
 - what is the definition of an 'Acquaintance'?
 - what units are the values measured in?
 - what do the codes mean?
 - only the publisher really knows

data may be available, but explanations are often hidden
 - in totally different documents
 - in non-machine-readable forms

second challenge: want to delve deeper
 - where is the crime defined in legislation?
 - which areas report the greatest levels?
 - how have levels changed over time, with changes in policy?

explicit links from within the data
implicit links to the same subjects
 - what other information is there about these crimes?
 - discovery through search, but with immediately usable results
combine these sources of data to create new visualisations



# Identifying things with URIs

## disambiguation & further data

Photo by Ivan Walsh http://www.flickr.com/photos/ivanwalsh/4244495312

RDF is a really good approach to tackle these challenges

RDF's only revolution, but the key one, is using URIs to name things, including properties and classes

identifying things with URIs does two really useful things
 - disambiguates, enabling joins with other data using same URI
    - mash-ups beyond mapping things on a Google Map
 - provides something at the end of the URI
    - extra information, explanation, context
    - in a basic entity-attribute-value model that enables combination without either up-front agreement or end-user jiggery-pokery

information can be referenced rather than copied
 - example is BBC's use of Wikipedia, MusicBrainz etc
 - huge opportunity for efficiencies within government, which excels at having five different versions of same code list managed by different people

identifying things with URIs is increasingly natural to web developers
 - using this principle for real-world things like schools or pillars and abstract things like classes and properties is less so, but even so

so why isn't it being used?
 - let's see what developers say

"The semantic web is given a rough raking by the syntactic web, and it is not impossible to see why when you first get taken down the SPARQL/RDF/ Ontology rabbit hole. **It is not great fun learning to develop with the semantic web today.**"

---

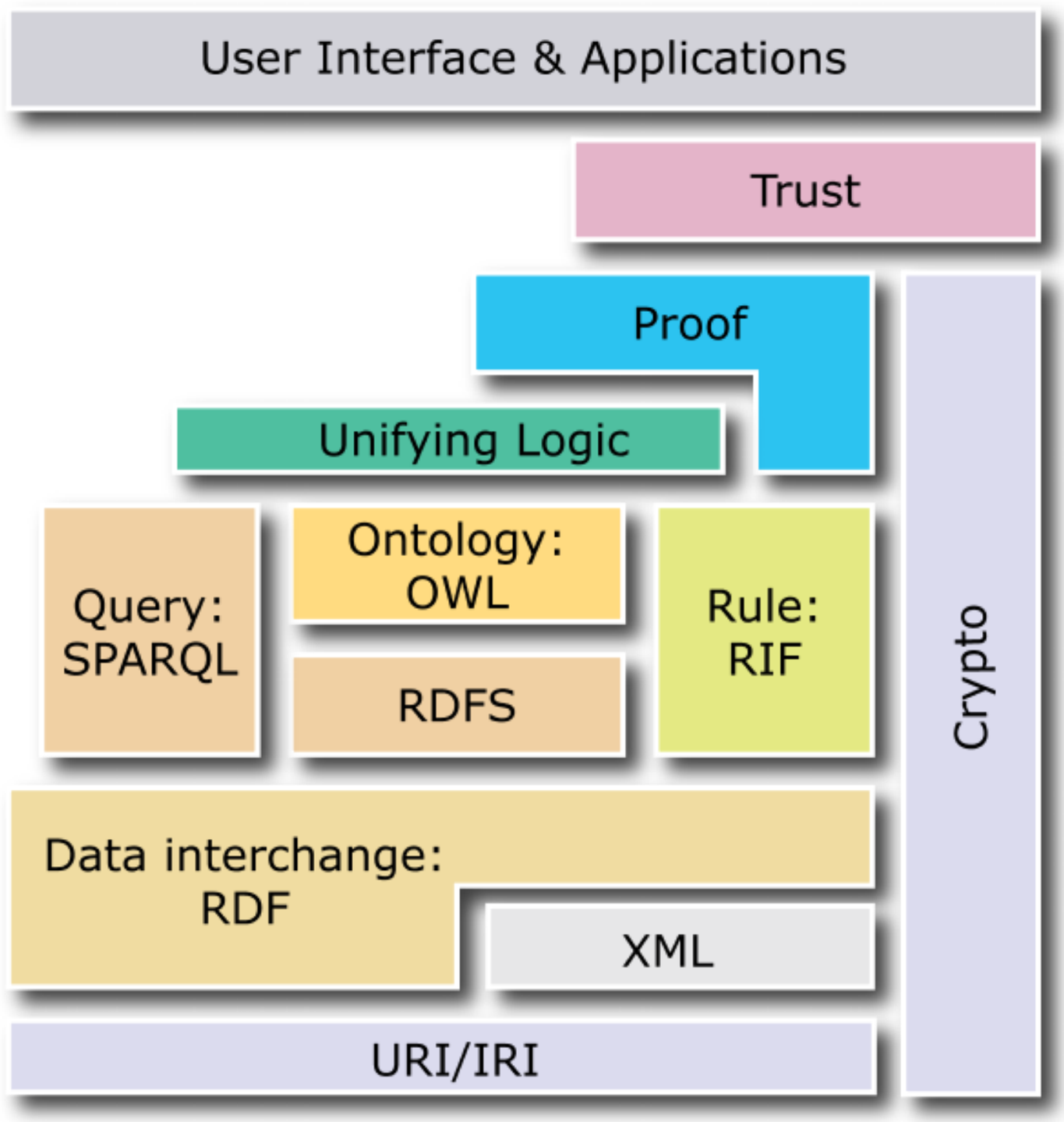
Daithi O Crualaoich
Linked Data at the Guardian

OPEN PLATFORM BLOG
Build applications with the Guardian

http://www.guardian.co.uk/open-platform/blog/linked-data-open-platform

delighted to see this a week or so ago as it's straight from the horse's mouth

over next three slides, put ourselves in developer's shoes




Do we need all this?

it cannot be this hard

developers encounter the Semantic Web as this stack of complex technologies
- even looks as though it might fall over on you!
- introduced to ugliest syntax for RDF, RDF/XML
- led to believe they need the most complex vocabulary for ontologies, OWL
- assume they need to provide access through triplestore & SPARQL endpoint

next couple of slides, explore this from two standpoints
- consumers of data, trying to build visualisations
- publishers of data, trying to get their data out there

# Consumption

make it fun to use

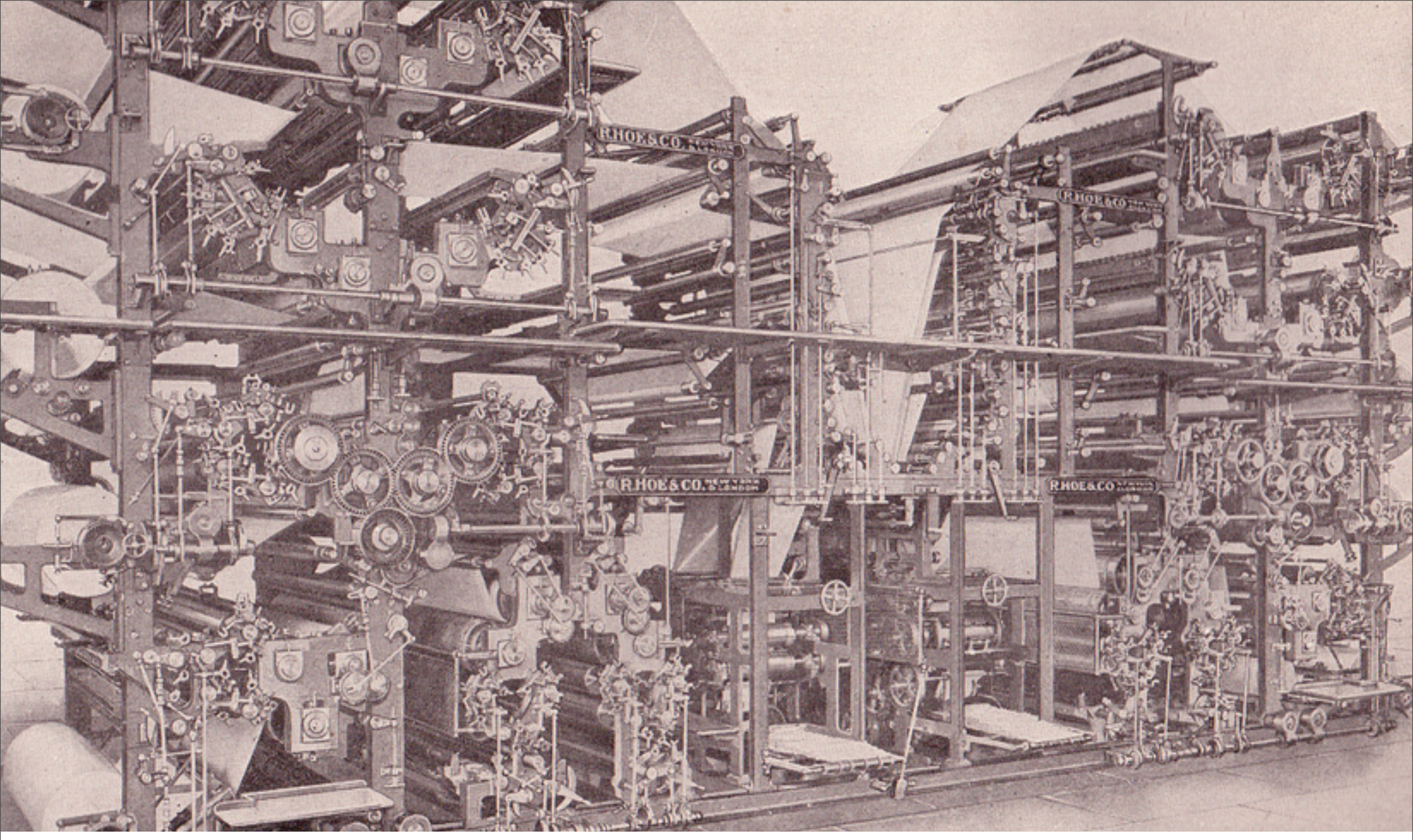
Photo by gamene http://www.flickr.com/photos/gamene/4074357954

standpoint of someone trying to use data to create a website or visualisation
 - is it fun and attractive?

compare with JSON
 - easy map into object structures
 - parsers on every platform
compare with early state of XML
 - standard APIs (SAX and DOM)
 - parsers on every platform
 - simple path languages for addressing (CSS and XPath)

RDF situation very different
multiple syntaxes
 - for triples (RDF/XML, N-triples, Turtle, N3, TriX, RDFa, RDF/JSON, RDFj)
 - quads (N-quads, TriG)
 - for SPARQL results
 - hard to know as a publisher what to target, as a consumer what to expect
no standard API mapping RDF structures into objects
no path language
 - SPARQL is as close as you get to standard querying
    - yet another language for developers to learn

no, it is not fun and attractive




# Publication

make it easy to create

Photo by Sue Clark http://www.flickr.com/photos/perpetualplum/3865517102

standpoint of someone trying to publish their data
 - is it painless and unlikely to go wrong?

library support for creating these formats is lacking for the same reason as parsing them is hard
 - generally generated directly from relational/object/XML structure rather than serialisation of an RDF graph
 - similar to creating XML using string concatenation
 - vast array of formats, not clear which are useful to target

dream of embedding RDF in HTML is often a nightmare
 - RDFa very difficult to get right
 - microdata verbose and doesn't cover the cases
 - Facebook RDFa shows the compromises that need to be made:
    - single vocabulary, fudge between literals and resources, snippable HTML in head of document

whole level of infrastructure around RDF publishing makes it even more complex
 - hash vs slash URIs
 - use of redirections
 - publication of vocabularies

publication is not painless; very likely to get negative feedback from data consumers for doing it wrong

---

these are the issues, what should we do


# Where next?

## what's needed

Photo by will ockenden http://www.flickr.com/photos/scissorhands33/3430164569

developers won't be persuaded by argument alone
- RDF needs to meet developers where they are
  - show fit with OO approach
  - tools and APIs that are easy and fun to use, slotting into their natural work
  - easy to follow recipes that avoid too much thought

focus on supporting end-user benefits developers can provide
for data consumers:
- clicking into visualised data to find out more
- easily incorporating data found elsewhere into their tools (eg adding extra columns in Freebase Gridworks/Google Refine)
- reach behind visualisation to trace data back to source

for data publishers:
- automatically increase the utility and value of their data because it's easier to combine
- ability to control context, give no excuse for misinterpretation
- ability to enhance their offering with others' data

network effect is really strong
- more people use linked data, easier it is to use, and more powerful its use becomes
- core hubs provide URIs that others can link to
  - don't have to invent own URIs for things
  - DBPedia, MusicBrainz, BBC, UK government etc
- core vocabularies publishers can reuse
  - FOAF, Dublin Core, SKOS etc
  - reduces need for invention, provides a focus for tools
- in government and commerce starting to be "everyone else is doing it" buzz

# Current work

what's happening

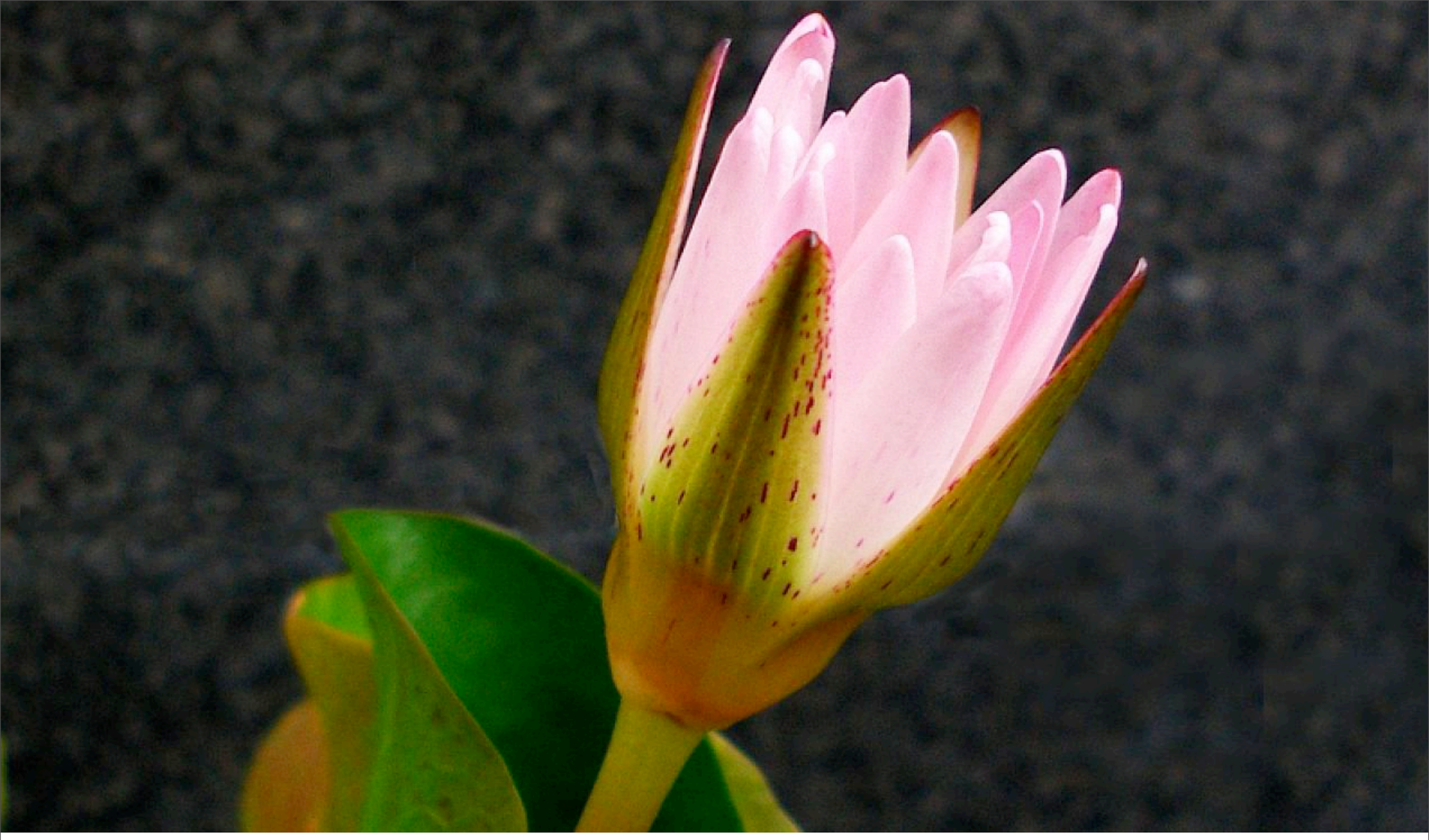
Photo by Jakob Montrasio http://www.flickr.com/photos/yakobusan/2436481628

promising current developments along these lines

linked data movement focusing on useful core technologies, not the stack
- creating guides and best practices
- need consumption to drive publication patterns
  - need to say, "if you do it like this, you and others can do Y better"
  - developers can make educated choices
- strong influence on RDF Next Steps workshop

on consumption
- generic RDF API embedded in RDFa API
- property paths in SPARQL 1.1: syntax for paths through data
- we need something like jQuery or nokogiri or hpricot
  - not something like DOM
- needs to be usable in contexts other than the browser

on publication
- RDB2RDF mappings: map relational to RDF structures
- are these going to be picked up for use in Ruby on Rails or Django?



# Fulfilling potential

## what should W3C do

Photo by tanakawho http://www.flickr.com/photos/28481088@N00/1344392695

tried to focus on meeting the challenges of the gradually blossoming web of data

linked data is a really powerful approach
- currently being avoided due to
  - stigma and misperception about RDF, views formed years ago
    - exacerbated by focus in SemWeb on logic and reasoning which are currently irrelevant for the web of data
  - lack of fun and easy to use implementations

what role can W3C play?
- hard to answer because biggest gaps are in implementations
- W3C blessing can help implementers make decisions
  - which formats to support, what API to implement
  - in a world of choices, this is vital
  - support profiling RDF and/or primer for usable core
  - provide a home for community standards
    - hard to trust vocabularies hosted on university domains
- W3C could extend services that help developers get started
  - syntax checking
  - validation against known vocabularies
  - visualisation of what they've produced
- more dramatically, W3C could sideline Semantic Web term
  - focus on web of data, of which linked data is a part

my biggest concern is that like every community, linked data community is too insular and may be deaf to wider developer concerns
- common human pattern to fall into self-reinforcing groups
- once you've adapted it's hard to see the gaps
- we only grow when we listen and change

W3C has a wide range of members
- you, not the existing SemWeb community, are the ones to listen to
- if you've avoided RDF for the last ten years, time to look again and make yourself heard

your involvement will help the web of data to bloom into something beautiful

# Comments?

Many thanks to John Sheridan, Thomas Roessler, Noah Mendelsohn, Dave Reynolds & Stuart Williams for comments on earlier version of this talk.