

eXist-db

@ London² XQuery Meetup



Saturday, 17th September 2011 @ The Wine Tun, London

Semantic Enrichment

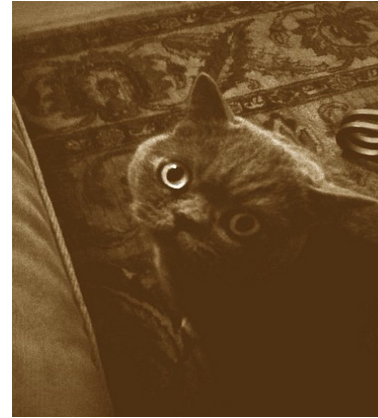
adam@exist-db.org



Why am I talking about this?

This was a bad idea...

“Intrigued” by the
Semantic Web



+



Lazy lunch with some old
friends from Local Gov

=

- My talk on two topics that I am new at:
- **Natural Language Processing (ish)**
 - **Semantic Web (things)**

What is Semantic Enrichment?



IMHO...

- Adding new Meaning to existing Content
 - Humans can easily identify within a page of text:
 - Names
 - Places
 - Events
 - References
 - For Computers this is un-natural – 101010
- The Semantic Web
 - Adding Meaning to the Web
 - Web of Data through Links (Linked Data)

- ...that Local Gov
 - Problems with existing proprietary Semantic search
 - Requirements
 - Classification of Content
 - Searchable Content (Full-text and/or Classification)
 - Semantic Enrichment (**Bonus Points!**)
- Surely there must be an Open Source solution?
 - So, Google it... for hours.
 - Components exist but some are complex and domain specific

- Classification
 - Against a custom poly-hierarchical vocabulary (**extra complexity**)
 - Bayesian Networks, Support Vector Machines etc.
 - ***Its complicated*** – *You have to train the computer!?!*
My next presentation ;-)
- Search
 - Just another Lucene index
 - ***EASY!***
- Semantic Enrichment
 - Build a spider, pass the results through some Web API's
 - **Should be easy.** *“Hey, what a great subject for a demo...”*

- Lots of free Web APIs for enriching **text** with Semantics
- RDFa is good for adding semantics to existing web pages
- ...Combine Semantic Enrichment and Semantic Web ideas
 - Get some **existing** content from the web
 - Extract the **text**
 - Pass it through some Semantic Engines
 - Post process the results from the Semantic Engines
 - Generate RDF and annotate existing content with additional markup and maybe some RDFa

The PAIN!



Lots of late nights down the code mine...

- All the Semantic Enrichment API's expect plain text
 - Only good during authoring*
- ...Well we can extract all text() nodes
 - Er... Surely we are discarding semantic meaning???
- Results are often string offsets to regions in the text
 - Relating string offsets back to text() nodes in a DOM, its really hard... (have code!)
- Splitting text() nodes to insert new element(s)
 - Really hard, can break markup, extents can span text(s)