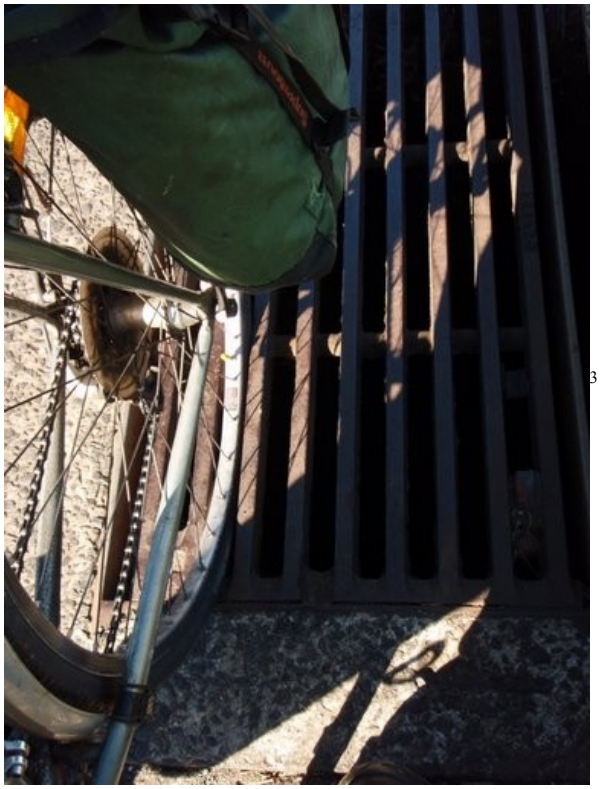
Adventures in Geocoding part 2: Embedding data points in documents

I have been thinking about how to start integrating more semantics into ICE[1] documents. This time I'll look at how you might embed geographical data in a document. This is only a preliminary look, but it's very promising so far.

A wrote a while ago about embedding metadata in pictures[2]. This time I look at how one might embed geographical data in a document.

I was tempted to do a dog-poo map of East Toowoomba showing how my hounds like to defecate as far as possible from a rubbish bin so I get to carry the re-used plastic bag further, but I'll spare you that and show you some thing else.

Take this cycle hazard for example. I have linked the picture it to a web album where you can see it in context. The caption has the location in the text so if you download the PDF you can find the location for yourself.

 [3]

One of many hazardous grates on Ruthven Street Toowoomba (-27.590334, 151.948166)

Or I could point out another dangerous place where the cycle lane disappears (-27.595667, 151.947174).

If everything is working correctly, you should see a map somewhere in this post (I'm still wavering about where to put it) showing those two points; and if you click on the little pins you will get the description for that point. I doubt it will work in places like Google Reader so click through to the post.

What I've developed here is actually not ICE specific. All I have done is adapt little bit of Javascript of Simon Willison's[4] to go through a page and look for HTML elements marked with the class attribute 'geo'. It's pretty dumb at the moment, and it relies on a convention that each location has an optional description followed by the coordinates in brackets. Only handles decimals, not degrees and minutes and would spit the dummy if you said 27.6045° S instead of -27.6045.

To use this in ICE I have to set up some javascript stuff to load everything in, install it in the blog server and so on, which took me a stupid amount of time, but in the documents themselves it couldn't be easier. I defined a new style called `i-geo` (i is for inline) and ICE automatically converts that to HTML spans with `class="geo"` when I generate the HTML.

By coincidence, there was a post this week from Roderic Page on mining PDFs for geographical data[5]. Great stuff. It's very like the work that Peter-Murray Rust's[6] group and others do with mining chemistry data from PDFs. But there are problems:

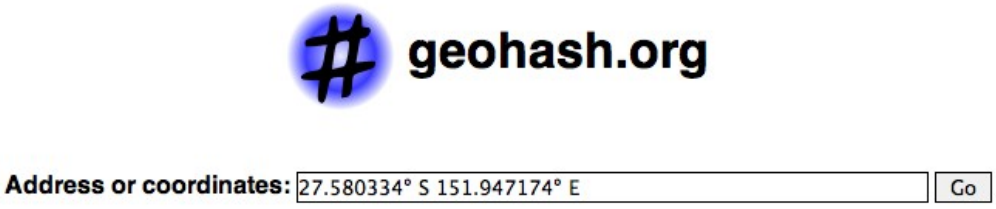
> The service uses a bunch of regular expressions to try and extract latitude and longitude pairs from the text (needless to say, there are nearly as many different ways to write a latitude and longitude as there are authors).
>
> http://iphylo.blogspot.com/2008/06/from-pdfs-to-google-earth.html[7]

What we want to do in ICE is provide authors with easy to use tools so they can unambiguously encode data and validate it **before** they hit the 'publish' button. One way we plan to do this is to adapt an application like Roderic's tool. In this case I'd point it at my document and it could tag all the coordinates I've got and normalize them to my preferred method or expressing coordinates, then mark them up in some way. Ultimately this will be more robust than these fragile after-the-fact scraping services. My document will be able to advertise its own meaningful content not cling to it jealously until it is exhumed by an application later on which has to pry it out of its cold dead PDF-fingers.

We're going to do something like this with the TheOREM project[8], too. It's in the work plan to run the OSCAR[9] chemistry-sniffer-outer application over documents and get it to mark all the bits of chemistry, as well as give its automatic sanity check; once that's done we can start pushing out chemistry with built-in semantics.

Now, I bet if Bruce D'Arcus[10] is reading this he'd be saying 'Use the new metadata support in ODF 1.2[11]' and I will investigate that. But, given the user base I deal with an OpenOffice.org only solution is not optimal – we also need a solution that will work for groups who use other tools such as Microsoft Word. The Style based microformat approach is one such interoperable mechanism. Styles work, but are bit tricky to apply. I like simple links even better.

In the geo-world geohash looks pretty cool. Geohash is an algorithm which can turn this:



Into this short URL: http://geohash.org/r7h51ehscv0g

I have set up my little Javascript experiment so that if I add a link it will automatically push a pin into my map. The Geohash algorithm is open, so I don't need the Geohash service to use it. I found an open source library easily enough[12], and the URL makes a perfectly good identifier in my opinion. Yeah yeah it's got the http protocol on the front but it's a unique string for a point on the earth and more importantly I can use it in pretty much any modern editor.

So I could use a simple link to point out a place where the road is in terrible condition [13]and have that point show up on my map. If you grab the PDF version of this page, you'll see that the links are all footnoted automatically so I don't have to type in coordinates or mess with styles. I just link. ICE has a nice feature where it can footnote all the links in my documents for the PDF version, too so the information is there in a usable form in print if we wanted to get really fancy it could decode the Geohash into a human readable forma for the print view.

What I'm thinking about now is a framework for semantic markup in word processors and beyond that takes into account all the prior art (smart-tags in Word for example) and the practical realities of mixed-application workgroups and a Microsoft-heavy world. I might try to put something together for the forthcoming e-Research in the Arts, Humanities and Cultural Heritage workshop[14]. We have a part-written paper about embedding metadata in documents lying around that may serve as a base.

1 http://ice.usq.edu.au/

2 http://ptsefton.com/2008/05/02/adventures-in-geocoding-part-1-the-toowoomba- bug-cycle-hazard-investigation-team-does-ruthven-street.htm

3 http://picasaweb.google.com/ptsefton/ToowoombaBUGHazardInspectionTeam/photo#5195619632523461826

4 http://24ways.org/2007/unobtrusively-mapping-microformats-with-jquery

5 http://iphylo.blogspot.com/2008/06/from-pdfs-to-google-earth. html

6 http://wwmm.ch.cam.ac.uk/blogs/murrayrust/

7 http://iphylo.blogspot.com/2008/06/from-pdfs-to-google-earth. html

8 http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/theorem-ice.aspx

9 http://www.rsc.org/Publishing/ReSourCe/AuthorGuidelines/AuthoringTools/ExperimentalDataChecker/getOSCAR.asp

10 http://community.muohio.edu/blogs/darcusb/

11 http://community.muohio.edu/blogs/darcusb/archives/2008/05/07/openoffice-30-beta-and-metadata

12 http://github.com/davetroy/geohash-js/tree/master

13 http://geohash.org/r7h51enkzthe

14 http://www.eresearch.edu.au/ahch-workshop