

Wordlink - current bugs and things to do

Caoimhín Ó Donnáil

SMO

2008-12-02

<http://www.claran.smo.uhi.ac.uk/wordlink/>

1. If you give it a link which should have a trailing slash, like
<http://www.smo.uhi.ac.uk/smo/dealbhan/arainn/ostaig/>
but omit the trailing slash, thus:
<http://www.smo.uhi.ac.uk/smo/dealbhan/arainn/ostaig>
it should still work, but it usually doesn't. It gets the base directory wrong and thus often fails to find pictures and stylesheets. This should be easy to sort and I have been looking for a solution, but haven't found it so far. I could probably do a temporary hack which would make it work for most situations.
2. In PHP, I feel that I should probably be using PEAR classes such as HTTP_Request2 and NET_URL, a modern technique which would probably give me better programming facilities. I have been reading up on this.
3. The links in the original document should work in the Wordlinked document (They do), and furthermore they should lead to documents which are themselves Wordlinked (They currently don't). This facility is available in the BBC VOCAB facility, in Google translate:
<http://translate.google.com/>
and in the "fun" Gaelic font program at:
<http://folkplanet.com/cgi-bin/mksc.r>
but it is not yet available in Wordlink. Apart from being important for the final production product, it would be very useful right now for testing, because we could browse to pages to test instead of having to copy and paste URLs.
4. In the case of words which are anchors for links in the original document, there is a conflict between the desire to link them to a dictionary, and the desire to make the original link work. The BBC Cymru VOCAB facility
<http://news.bbc.co.uk/welsh/>
manages to achieve both. The original link works, but a JavaScript popup gives (as well as a basic translation) the option to follow a link to a dictionary. I'd like to do the same, at least as an option.
5. For dictionaries which we can get possession of (like the Stòr-dàta Briathrachais Gàidhlig), I'd like to give Wordlink the option of displaying the dictionary definition immediately, as a JavaScript popup as the BBC VOCAB does, rather than the user having to click and switch to a dictionary page and then click back.
6. I'd like to put more options in Wordlink's navigation frame - The ones I mentioned above, and also options to change the background-color and other style of linked words.
7. I could easily throw in more languages, apart from the current Gaelic and Danish, and I probably should do, so that more people can test it. What I want in the end, though, is a database of dictionaries, giving the source and target languages, and the strings and

methodologies required to link to the dictionary. It would then be easy to add more many more languages and dictionaries, and Wordlink could dynamically (using Javascript) change the available target languages as the user changed the source language, and dynamically change the available dictionaries as the user changed the target language. I have made a bit of a start on this database of dictionaries.

8. The Danish dictionary which I am currently linking to appears to me to have been “stolen”, although it is openly available on the Internet. OK for test purposes, but we need to find another one.
9. Gaelic, as Gordon points out, badly needs lemmatisation (although I have done lenition removal by a hack). The BBC VOCAB facility at:
<http://www.bbc.co.uk/scotland/alba/naidheachdan/>
is in very much the same position. I have been given kindly given a “starter” set of Gaelic wordforms by a friend, and will turn this into a “starter” lexical database for Gaelic, which will make a big difference to this problem. I recognise, though, that I can’t devote too much time out of this POOLS-T project to Gaelic lemmatization. Lemmatization will be a question which will need addressing for all languages, although some will need it more, some less.
10. Wordlink is currently using “frames”. Frames are a bright idea from the early days of the Internet (mid 1990s), which quickly caught on and have now very much gone out of fashion, being practically ignored by new developments in Web standardization. I never liked them myself, but Wordlink might be the perfect application for them. However, I am currently reading up on “Iframes”, in case that is another possible methodology, and am also thinking about the possibility of displaying the document in an ordinary HTML `<div>..</div>`.
11. I am currently blithely assuming that the original document encoded in Unicode/utf-8. This will be the case for all Web documents in a few years from now, but might currently be true for less than half of all Web documents. I haven’t even begun to test whether Wordlink works for pages in other encodings, or to see what coding conversion could be done if it doesn’t.
12. There are lots and lots of other technologies used by Web pages which might potentially mess up the workings of Wordlink: XML; JavaScript; frames; Iframes; entities; Flash, etc, etc. I haven’t even begun to test what happens with these.
13. I changed the name to “Wordlink” to be shorter, but if anyone thinks it should be “Wordlinker”, let me know.