Footnotes, comments, bookmarks, and marginalia on the Web - **A W3C Workshop on Annotations**

2 April 2014, San Francisco, California

Precision phrase linking and pulling – ispantu[*]

Gavin Brelstaff
CRS4 Sardinia Italy
gjb @ crs4.it

Francesca Chessa
University of Sassari, Italy
fch @ uniss.it

POSITION PAPER

Introduction: *Precision semantic linking and pulling:*

The semantic web can be interpreted as a web of semantic units – or ideas. An idea can occasionally be found succinctly expressed in a phrase on a remote web-page. To make sense of the semantic web we ought to be able to link precisely to such a text-phrase, plucking it from its surroundings and re-presenting it, while intrinsically attributing its derivation: its semantic context. A new page could be built from a mosaic of such verbatim citations pulled *in situ* from over the web. For this to be of enduring value the integrity of phrases once cited should be verifiable later. All this becomes less of an academic exercise when one considers the multilingual field – where, for challenging translations, re-tracing the precise source-language context often assists a broader comprehension.

Multilingual context

Indeed the web-page can act as a natural fabric for visualising phrase-based parallel-text translation in an adjacent column format – as dynamically demonstrated at recent W3C *MultilingualWeb* Workshops[†]. There we used client-side scripting to gain access to text contained within phrases found inside our web-page or on an external page: Once a page is loaded a jQuery/CSS selector approach – e.g: `$('span#phrase2').text()` – delivers the content of a given phrase for semantic alignment. Our potential for building a multilingual semantic network, however, is severely impeded by the *same-origin policy* that prevents scripted access across external domains – even for text!

The basic idea: *ispan*

If it is possible to use an `iframe` to embed a page from another domain in your site why can't a text fragment from that page be similarly embedded in, say, an `ispan`? – i.e. a `span` element extended to interpret a `src` attribute like an `iframe` does. Here we are simply talking about scraping text from one web-page onto another – not subverting another domain's layout, content, scripting or form submission. Ensuring that only text is accessible would be a sobre user-agent duty, along with its other same-origin policy responsibilities.

How might this work? Extending the existing HTML `span` element would suffice. The new `src` attribute would simply be an URI addressing an anchor in the other page using the # suffix to identify the desired fragment, e.g. in our page we would write:

```
<span id='#my-phrase'
src='http://elsewhere.com/theirpage.html#their-phrase'>replace
this</span>
```

and on the remote page a `span` element containing the desired text fragment would have the matching `id` attribute:

```
<span id='their-phrase'>Is this what you want to read?</span>
```

Clearly an opt-in approach is needed here since it would be dangerous to allow the arbitrary scraping of private content (like credit card details) from one domain to another. Opting-in could be done by attaching an attribute `share` to any element in the page that nests within it the spans to be shared as text. E.g.

```
<article share='text'>
   . . .
   <span id='their-phrase'>Is this what you want to read?</span>
   . . .
</article>
```

User-agent requirement: *access, attribution and integrity*

It remains for the user-agent to act on any such addressability and isolate the content from a source page in an appropriate manner. What is written on the right-hand side of the # sign has always been the responsibility of the user-agent – which dictates the entire source page gets pulled even to extract one tiny phrase. A mosaic made from numerous source pages may thus need to adopt a link-prefetching strategy like that made available in HTML5 – in any case, bandwidth necessarily gets consumed in a live semantic network. The user-agent ought also invoke an agreed graphical link-emblem to express attribution of the source – derived from the `src` attribute described above. That emblem might change its form/colour to indicate an offline situation where a cited phrase cannot be accessed live. Which leads to a first use-case

Use-cases 1: simple mirroring

Here our page is authored with the remote phrase already pasted into our `span`. The `src` attribute simply points to the fragment of the original page that supplied the phrase. If on-line, the user-agent can verify the integrity of the copy or check if the original has changed.

```
<span id='#my-phrase' src='http://elsewhere.com/theirpage.html#their-
phrase'>Is this what you want to read?</span>
```

The integrity of a source phrase can be compromised in three distinct ways: (a) the web-page is not found (HTTP status code 403), (b) the fragment-id is not found in the web-page, and (c) the content addressed by the fragment-id has been modified. The latter case, may be revealed by (i) the HTTP header `Last-Modified` field, or (ii) comparison with the pre-copied span content above. One might even consider using cryptographic hashes, in place of verbatim text, in order to check integrity but only when dealing with long phrases.

Use-cases 2: Multilingual*Web*

Here the idea is to make simultaneously available both the translated and original-language phrase – so that their display might be toggled: Our `span` element houses the translated phrase and its `src` attribute is set up to pull the original-language phrase if and when needed. Again to cater with off-line scenarios, the verbatim original-language phrase could be pre-stored in an agreed attribute (`equiv`) e.g. as follows

```
<span id='#my-phrase' equiv='la:cogito ergo sum'
src='http://elsewhere.com/theirpage.html#theirphrase'>I think
therefore I am</span>
```

The value of the `equiv` attribute here gets prefixed (in a name-space style) with the standard language code: i.e.. `la:` for Latin. The remote page could contain something like this:

```
<article xml:lang='la' share='text'>
   . . .
   <span id='their-phrase'>cogito ergo sum </span>
   . . .
</article>
```

Here, the integrity of the source phrase would be checked using the value of the `equiv`, when online. There curious scholar might like to navigate to the point in the linked article to better glean the semantic context of the source under translation.

Further extensions?

Could the basic scheme be extended to allow rich-text fragments (**`share='rich-text'`**)? – italics, for example, often convey intrinsic meaning. Some way of normalising white-space may be need to be specified.

Could remote fragments be identified using CSS style selectors? e.g. something like: `http://elsewhere.com/theirpage.html`**`$('span.myclass').`**

For collaborative multilingual sites might the fragment-identifier `id` actually be specified as verbatim text? – thus obviating the need for the `equiv` attribute – In that case, navigating to an anchor would become, for the user-agent, almost akin to searching for a phrase within the pulled page.

NOTES

* ***Ispantu*** in the Sardinian language translates to *amazement, wonder, miracle*.

† **MultiLingual*Web*** initiative http://www.multilingualweb.eu/ – our videos here:
http://www.multilingualweb.eu/documents/rome-workshop/rome-program#setting
http://videolectures.net/w3cworkshop2011_brelstaff_interactive/#video_player_ajax

Resources consulted

RFC 5147, 2008. http://tools.ietf.org/html/rfc5147
URI Fragment Identifiers for the text/plain Media Type
which proposed ways to address text fragments in plain-text pages: e.g.
`http://example.com/text.txt#line=10,20`

HTML5 http://www.w3.org/TR/html5/links.html
which allows anchors within `article` elements to be given the `rel` attribute `"bookmark"` – so identify that `article` as a precise fragment in the web page. Also links may be marked for prefetching with *the `rel` attribute `"prefetch"`*

Cross-Origin Resource Sharing - W3C Recommendation 2014
http://www.w3.org/TR/cors/ *allows mutually consenting web-pages to follow a formal HTTP protocol to share their contents.*

HTML5 Web Messaging Candidate Recommendation 2012
http://www.w3.org/TR/webmessaging/ *proposes the sharing of text information via an agreed, scripted, messaging protocol. For our purposes, adopting such an elaborate protocol simply to share text seems like a significant overkill.*

JSON-P http://json-p.org/
a protocol for lightweight data-interchange via asynchronous data transfer over the internet. This carries javascript objects not their html instantiations.

jQuery Selectors http://api.jquery.com/category/selectors/
simplifies access to HTML DOM nodes on scripted web-pages.

We thank the generous support of the *Fondazione Banco di Sardegna* and the *Regione Autonoma della Sardegna*. Thank you also to Prof. Nicola Tanda.