Enabling scholarly publishers to sync annotations among syndicated copies of articles

The Problem

The article *HDAC1 and HDAC2 control the specification of neural crest cells into peripheral glia*, is published in the Journal of Neuroscience at this URL http://jneurosci.org/content/34/17/6112

It has this DOI: 10.1523/JNEUROSCI.5212-13.2014 (which resolves to the above URL)

The article is also syndicated to:

Pub Med URL: https://www.ncbi.nlm.nih.gov/pubmed/?term=PMID%3A+24760871
Pub Med Central (US) URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3996228/
Europe PMC URL: http://europepmc.org/abstract/MED/24760871

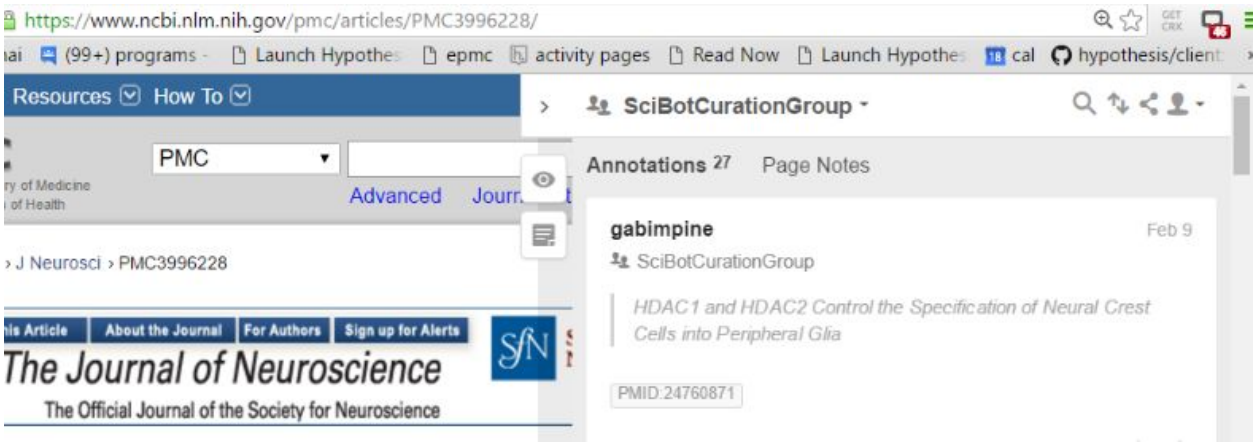
Annotations made against any of these URLs should coalesce with annotaions to any other. Instead here is the situation:

(annotation count, url)

(46, 'https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3996228/')
(46, 'doi:10.1523/JNEUROSCI.5212-13.2014')

(0, 'http://jneurosci.org/content/34/17/6112')
(0, 'https://www.ncbi.nlm.nih.gov/pubmed/?term=PMID%3A+24760871')
(0, 'http://europepmc.org/abstract/MED/24760871')

In this screenshot we see what we expect (Note 1: The badge reports 46 which is the sum of the 27 annotations noted in the sidebar and 19 replies to those annotations. Note 2: All annotations are in the SciBotCuration group.)



In this screenshot we don't:

Background

In 2009, Google, Microsoft, and Yahoo announced support for a new link element, rel="canonical" which, Google's Matt Cutts wrote at the time, enabled publishers "to clean up duplicate urls on sites." In 2012, https://tools.ietf.org/html/rfc6596 formalized the idea. It was mainly intended to coalesce varying URL patterns witrhin individual sites. E-commerce sites, for example, often provide multiple paths to the same page. To improve search engine optimization, this approach enabled them to pick a single cannonical URL and point crawlers at that from all the variants.

A less common use case for rel="canonical" enables this coalescence to happen across domains. When that mechanism is used, Hypothesis creates equivalences among pages that point to the same canonical URL.

Possible Solution 1: Syndicators use rel=”canonical” to point to the original journal

In this example, if we consider http://jneurosci.org/content/34/17/6112 to be the canonical URL for that article, it would be possible for PubMed, PubMed Central, and Europe PMC to include the following in the HEAD element of their HTML pages:

```
<link rel=”canonical” href=”http://jneurosci.org/content/34/17/6112”>
```

Hypothesis would then coalesce annotations among http://jneurosci.org/content/34/17/6112 and the others.

Problems with this approach

- Per RFC6596, there should be only one canonical link relation for a resource. If used for this purpose it would not be available to coalesce multiple paths within a site
- Other?

Possible Solution 2: Hypothesis coalesces syndicated pages that point to a common DOI

The reason 46 annotations are found in the table above, for https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3996228/, is that it was the URL against which thos annotations were made. No rel=”canonical” link in that page tells Hypothesis to associate it with http://jneurosci.org/content/34/17/6112. And no such link in the PubMed or PubMed Central pages carries that association transitively to them.

But there is another Hypothesis query, a uri: query for **doi:10.1523/JNEUROSCI.5212-13.2014,** that finds the 46 annotations. Why? The annotated page, www.ncbi.nlm.nih.gov/pmc/articles/PMC3996228/, includes this metadata:

```
<meta name="citation_doi" content="10.1523/JNEUROSCI.5212-13.2014" />
```

That same declaration is found in: jneurosci.org/content/34/17/6112 and europepmc.org/abstract/MED/24760871 (which also includes an alternate way to cite the DOI, `<meta name="dc:identifier" content="http://dx.doi.org/10.1523/JNEUROSCI.5212-13.2014"/>`)

(The citation_doi is not, however, found in www.ncbi.nlm.nih.gov/pubmed/?term=PMID%3A+24760871, instead the DOI is referenced like so: `<meta name="description" content="J Neurosci. 2014 Apr`

23;34(17):6112-22. doi: 10.1523/JNEUROSCI.5212-13.2014. Research Support, N.I.H., Extramural; Research Support, Non-U.S. Gov't" />)

So, Hypothesis could coalesce articles that share a common DOI. In this example annotations made against any of the above URLs except www.ncbi.nlm.nih.gov/pubmed/?term=PMID%3A+24760871 would coalesce.

Problems with this approach

- ?