Inspiratron.org - Natural language processing, machine learning and cybersecurity

Moment when my idea became a web standard

by Nikola Miloševi? - Thursday, June 15, 2017

https://inspiratron.org/blog/2017/06/15/moment-idea-became-web-standard/

This is the story how one schema I worked on as a side project suddenly found its place in W3C recommendation.

In November 2015, I went with my supervisor to Japan. In small cities of Mishima and Ito, about 1 hour train ride from Tokyo was held Biomedical Linked Annotation Hackathon (BLAH2) to which my supervisor was invited. He could not stay for the whole period, so he offered me to go, which I accepted. The event was organised by Japanese Database Center for Life Sciences (DBCLS).

On the first day was the conference, where people were presenting their work mainly on annotating biomedical literature. My PhD was related, kind of similar topic, it was about information extraction from tables in biomedical literature. However, I had not presented, only my supervisor and some other people who often presented works of the research groups that they lead. There was about 25-30 people, but quite involved people in the topic.

On the second day we moved to Ito to have 4 day hackathon which will advance biomedical annotations. As annotation framework was mainly used PubAnnotations that was annotation schema developed and maintained by DBCLS. However, this annotation schema, as majority annotation schemas at that time were developed for annotating text. Annotation schema was ignoring the structured part of the document, even though at that time had capability of splitting articles from PMC to sections. 5 years before the event was also started JSON-LD, so couple of people like Lars Juhl Jensen and Sampo Pyysalo worked on that.

I have decided to design schema extension for PubAnnotations that will allow annotation of XML documents and the structure inside the documents, including tables. The main issue with annotating documents in XML is storing the location of the annotation. No more can be used string span and how many characters away from the beginning is the annotated word or phrase. So I searched for the technology that will allow me locate annotations in the structure of XML. I found technology called XPath that is able to locate path to the given XML element. Also XPath supports substrings inside the located element, which was perfect way to show annotation tool where the annotation is located. The types of annotations I kept from PubAnnotations with denotations, relations and modifications and format remained the same (json). I run couple of discussion sessions, especially with Jin-Dong Kim, who is the director of DBCLS and the main developer behind PubAnnotations. We mainly agreed that it was a good idea and Jin-Dong had even thoughts about including it at some point in PubAnnotations. However, at this hackathon I had no time to make an annotation tool. Since short time was left, I developed couple of conversions tools for different annotations formats to PubAnnotation format, such as MetaMap2PubAnnotation (for conversion of UMLS annotations generated by MetaMap into PubAnnotation)and GATE2PubAnnotations (for conversion of annotations created by GATE into PubAnnotations).

A year later (summer 2016) BLAHmuc was held in Munich, Germany. I went with the proposal to create an annotation tool based on the schema I developed year earlier in Japan. My project proposal can be found on this location. I worked on it for a week, finding a method to allow tool to annotate XML based on overlays in javascript. In these 4 days of hackathon, I managed to put together proof-of-concept web application that can be found on GitHub. Unfortunately, it was not really a product that could be used for real scale annotations. Again, I had couple of discussion rounds with Jin-Dong about my schema and how it can be integrated. We even talked about some scholarship schemes in Japan that I can apply and implement it myself (unfortunately it so far didn't happened as I will be staying at Manchester University for a bit longer).

Couple of days ago, I was searching for some web annotators and doing my literature review. Suddenly, I ended up on this page: https://www.w3.org/TR/annotation-model/. The W3C, or WWW Council, a body making web standards (calling them recommendations), released on 27th February 2017 web annotation data model that uses XPath for locating annotations in XML/HTML documents. The idea, I had 2 years before became a part of web standard. Unfortunately, I was not referenced, and since it is quite a logical way for solving the problem, I can believe that they made it independently. Searching a bit more, I found that similar idea was developed also at one group at Postdam University, with annotation schema called PAULA XML. Their documentation is from 2013 and it is using XPointers in order to locate annotated elements. W3C is using JSON-LD that they proposed for annotations, instead of our PubAnnotations, however the way of locating annotated elements is identical. There are 3 ways how to make selection in W3C, however, XPath selector way is the one I proposed 2 years earlier. I unfortunately can't claim that they copied idea, it even makes me excited about it. It is also interesting how ideas can emerge on different places.

In order to prove here is how XPath selector looks like:

EXAMPLE 22: XPath Selector

```
{
  "@context": "http://www.w3.org/ns/anno.jsonld",
  "id": "http://example.org/anno22",
  "type": "Annotation",
  "body": "http://example.org/note1",
  "target": {
    "source": "http://example.org/page1.html",
    "selector": {
      "type": "XPathSelector",
      "value": "/html/body/p[2]/table/tr[2]/td[3]/span"
    }
  }
}
```

And here is how my proposal for PubAnnotation looked like:

```
{
   "xml": "<table>
              <tr><td>parameter</td><td>number</td></tr>
              <tr><td>male/famale</td><td>15/18</td></tr>
           </table>",
   "denotations": [
      {"id": "T1", "xpath": "/table/tr[1]/td[1]", "obj": "Header"},
      {"id": "T2", "xpath": "/table/tr[1]/td[1]", "obj": "Stub"},
      {"id": "T3","xpath": "/table/tr[1]/td[2]", "obj": "Header"},
      {"id": "T4","xpath": "/table/tr[2]/td[1]", "obj": "Stub"},
      {"id": "T5", "xpath": "/table/tr[2]/td[2]", "obj": "Data"},
      {"id" : "T6","xpath": "substring(/table/tr[1]/td[1],2,5)",
"obj": "substringEx"}
   ],
   "relations": [
      {"id": "R1", "subj": "T4", "pred": "dataOfHeader", "obj": "T1"},
      {"id": "R2", "subj": "T5", "pred": "dataOfHeader", "obj": "T3"},
      {"id": "R3", "subj": "T5", "pred": "dataOfStub", "obj": "T4"},
   ]
}
```

They are putting xpath in value, while I proposed path property that will be filled with xpath.

To me this is quite amazingly interesting. I would actually like to know whether there was another stream of development, not mentioned

here (so not over BLAH events and University of Postdam) and how about came this idea and what is the genesis. This is my part being made public, I would like to hear from them as well.

Now there is also quite nice tool that utilises this web annotation data model, called hypothes.is being a web service with API and Chrome plugin, therefore it is much easier to make web annotations and annotations of structured documents than it used to be a number of years ago, when I started PhD and when there was no annotations for semi-structured documents, such as web pages, and elements in them like tables.
