WebAnnotator, an Annotation Tool for Web Pages

Xavier TANNIER
LIMSI-CNRS, Univ. Paris-Sud, Orsay, France
xavier.tannier@limsi.fr

Manually Annotating Web Pages

Needs for manually annotating Web pages are many:

- Text tagging
  e.g. named entities
- Image tagging
  e.g. image retrieval
- Web page cleaning
  e.g. ad detection, metadata, blog detection, tables...
- etc.

WebAnnotator Objectives

❶ Annotating online pages
and not having to store and clean them before.

❷ Maintaining visual rendering of HTML
so that annotation is made easier and closer to real user experience

→ **Firefox add-on**

❸ Allow annotation of any element in the page
not only text but also images, menus, etc

❹ Allow both human- and machine-readable- saving formats
even if the original page is ill-formed or if annotations overlap HTML tags

https://addons.mozilla.org/en-US/firefox/addon/webannotator/

Overview

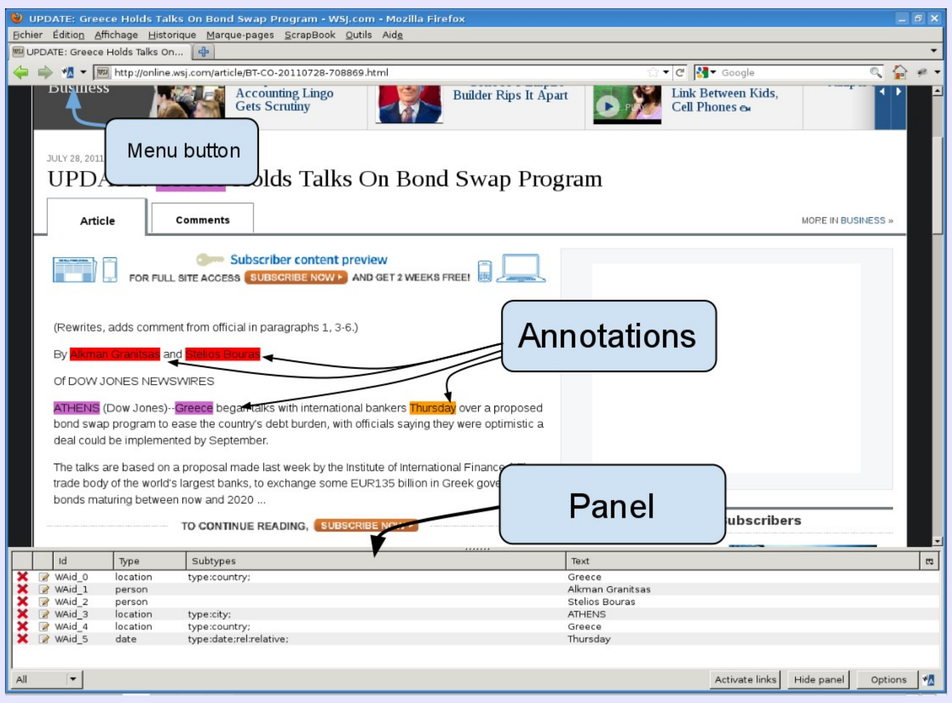
Why a Firefox extension?

- Firefox is commonly used, and people are used to install extensions
- Firefox is a web browser and there is no chance we can guarantee the visual rendering of HTML better than it does
- Everything that can be selected in Firefox can be annotated

↳ Needs ❶ ❷ and ❸ are naturally fulfilled.

How does WebAnnotator work?

- Users can specify their own annotation schema (DTD)
- Both online and offline pages can be annotated
- Annotations can be saved (HTML with highlighted segments) or exported (machine-readable format)



- A button and a panel are added to the Firefox view
- Annotations are made directly on the Web page
- The bottom panel records all annotated segments

Creating an Annotation Schema

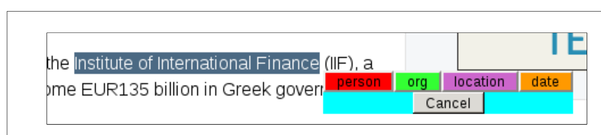
User-defined DTD (inspired from Callisto)

```
<!-- Four high-level annotations types : person, org, location, date -->
<!ELEMENT person (#PCDATA)>
<!ELEMENT org (#PCDATA)>
<!ELEMENT location (#PCDATA)>
  <!-- Attributes for locations, no default -->
  <!ATTLIST location type (river|mountain|city|country) #IMPLIED>
<!ELEMENT date (#PCDATA)>
  <!-- Attributes for location              -->
  <!ATTLIST date type (date|time|duration) #REQUIRED
                 rel (absolute|relative) absolute
                 value CDATA #IMPLIED>
```

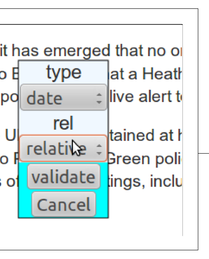
Allowed types are `person`, `org`, `location` and `date`. Type `location` has an optional attribute `type` that can take the values *river*, *mountain*, *city* or *country*. Type `date` has two required types: `type` and `rel`. This latest has a default value *absolute*. The optional `subtype` value is a free-text attribute.

Annotating Pages

Select-and-choose

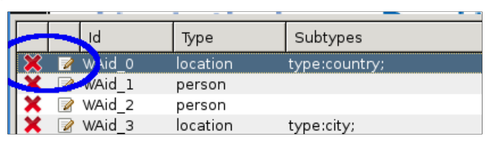


When selecting a segment, a small rectangle pops up and the user can choose the annotation type.



If this type contains specific attributes (as specified by the loaded DTD), the user can choose their values.

Two ways and modifying annotations: near the highlighted segment (left) or from the bottom panel (right)



Saving and Exporting

- Need to avoid element overlapping
  otherwise HTML is no longer valid (or even more invalid)
- Other systems propose separated, stand-off markup
  which we do not want. Annotations are just another markup of the file and can be strongly related to rendering and context.
- We must be able to continue our annotation on Firefox after saving

↳ Two formats: "save" and "export"

Original HTML code

```
Annotation schemas can be specified to WebAnnotator by <b>importing a DTD</b>
```

HTML rendering

Annotation schemas can be specified to WebAnnotator by **importing a DTD**

HTML rendering of an annotation on *"by importing"*

(overlap)

Annotation schemas can be specified to WebAnnotator **by importing a DTD**

Save

```
Annotation schemas can be specified to WebAnnotator
<span wa-id="1" wa-type="PP" class="WebAnnotator_PP">
  By
</span>
<b>
  <span wa-id="1" wa-type="PP" class="WebAnnotator_PP">
    Importing
  </span>
  a DTD
</b>
```

- *Keep the exact same rendering*
- *Allows to carry on your annotation task*

Export

```
Annotation schemas can be specified to WebAnnotator
<WA_start wa-id="1" wa-type="PP" />
  By <b>Importing <WA_end wa-id="1" />
  a DTD </b>
```

- *Replaces HTML span tags by empty XML elements*
- *Automatic processing is easier*
- *Results in valid XML (if the Web page is valid XHTML...)*