Automatic Text Analysis with oXygen

XML Prague 2014

oXygen Users Meetup

Felix Sasaki

DFKI / W3C Fellow

Aim

- Identifying concepts in text is useful
  - Disambiguation as a preparation for translation, search engine optimization, various type of content analytics applications
- Issues
  - manual annotation takes too long
  - automatic annotation is error prone
- Solution: enable automatic annotation in oXygen and allow users easily to edit results
- All files needed + installation steps (for oXygen 15.1 and oXygen 15.2) are at

http://www.w3.org/International/its/wiki/Its2-in-oXygen

Steps

- 1) Customize your content type(s) so that text analysis markup is allowed
  - Here done for DocBook 5
- 2) Create an oXygen action to generate the text analysis markup
  - Call of one external annotation service
- 3) Create a WYSIWYG environment for easy correction

1) Customize document type

- Allow “Internationalization Tag Set (ITS) 2.0” markup ”Text Analysis” http://www.w3.org/TR/its20/. Example:

```
<para>Welcome to the
<phrase
Its:taClassRef="http://nerd.eurecom.fr/ontology#Location"
Its:taIdentRef="http://dbpedia.org/resource/Prague">
city of Prague</phrase>!</para>
```

1) Customize document type

- Customization provided by Jirka Kosek, see

http://xmlguru.cz/2013/05/docbook-and-its2

- Add dbits.rng to

oxygen/frameworks/docbook/5.0/rng

- Set DocBook 5.0 type to use the schema

${framework}/5.0/rng/dbits.rng

2) Create an oXygen action to generate text analysis markup

- Here: XSLT stylesheet that
  1. Extracts text to be annotated
  2. Sends the text to RESTful online service, here DBpedia spotlight https://github.com/dbpedia-spotlight/
  3. Integrates annotations back into original content. IMPORTANT: try to preserve existing markup as much as possible*

Algorithms relevant for step 3 are defined at

http://www.w3.org/TR/its20/#conversion-to-nif

http://www.w3.org/TR/its20/#nif-backconversion

* Won't work if annotations would lead to overlapping markup

3) Create a WYSIWYG environment for easy correction

- Use oXygen customized CSS stylesheets
- A big thanks to George Bina for introducing me to this cool feature ☺
- Add the stylesheet as an option for DocBook 5 CSS settings

DEMO

Steps (again)

- 1) Customize your content type(s) so that text analysis markup is allowed
  - Here done for DocBook 5
- 2) Create an oXygen action to generate the text analysis markup
  - Call of one external annotation service
- 3) Create a WYSIWYG environment for easy correction

What next?

- 1) Customize your content type(s) so that text analysis markup is allowed
  - Here done for DocBook 5 more to come
- 2) Create an oXygen action to generate the text analysis markup
  - Call of one external annotation service more to come
- 3) Create a WYSIWYG environment for easy correction could be easiser ☺

Provide feedback and contribute

- "What are your use cases for content analytics?"
  http://tinyurl.com/co-an-survey
- That may be related to XML – or not
- Also: gather feedback in free text form: please join me at http://tinyurl.com/ca-gdocs
- Join the W3C LD4LT community group to discuss these topics
  http://www.w3.org/community/ld4lt/

Automatic Text Analysis with oXygen

XML Prague 2014

oXygen Users Meetup

Felix Sasaki

DFKI / W3C Fellow