BibTeX Bibliography Index Maker: Meeting Notes

Ramon Xuriguera

25-03-2010

1 Comentar informe

2 Wrapper induction

Tenir com a mínim dos exemples de la pàgina etiquetats. Per nosaltres, les etiquetes seran aquells trossos d'informació que ens interessa extreure.

- Posició relativa dins del document:
  Calculem la posició *bottom-up*, des de les fulles i fins on sigui necessari. Si no necessitem arribar a l'arrel per identificar l'element, no ho farem.

```
[(u'table', {u'width': u'100%'}, 7), (u'tr', {}, 0)]
```

  A l'hora de crear el *path* es podria mirar d'utilitzar el número de *sibling* per desambiguar. (Ara per ara, aquest número es posa a la llista, però no es té en compte per determinar si l'element passa a ser únic). Això permetria tenir rutes més curtes i poder estalviar passos a l'hora d'obtenir l'element.

  Utilitzarem tant etiquetes com atributs. Tots els atributs, excepte aquells que sabem que poden variar entre document i document del mateix lloc web, per `onclick` o `href`. Els que segurament ens seran més útils són `id` i `class` ja que no variaran entre document i document del mateix lloc web i són molt comuns ja que avui en dia gairebé tots els llocs web utilitzen CSS.

- Posició relativa dins de l'element:
  De la mateixa manera que WHISK (Stephen Soderland 1999), si el contingut de l'element no només conté la peça d'informació que estem buscant, l'haurem d'extreure utilitzant expressions regulars.

  Podem crear aquestes expressions regulars mirant la resta de contingut de l'etiqueta. Comencem agafant el fragment d'informació que ens interessa i mirem si està acompanyat d'altres caràcters. Si és així, agafem els caràcters del voltant. Finalment creem l'expressió regular. Per exemple, pels següents camps:

  - Info: `2007`
    Text complet: `Year of Publication: 2007`
    Regex: `p;(.*)`

- Info: `0925-2312`
  Text complet: `ISSN:0925-2312`
  Regex: `N:(.*)`

- Info: `2668-2678`
  Text complet: `Pages:  2668-2678  `
  Regex: `:  (.*)&n`

Tot i funcionar la majoria de vegades, ens podem trobar en que els caràcters que envolten la informació que ens interessa també variïn entre pàgina i pàgina. Per això cal comparar l'expressió regular que hem obtingut amb la resta d'exemples disponibles.

Una altra manera de fer això podria ser agafant directament el text del mateix element corresponent a diferents exemples i comparant-los. Així també podríem saber si el text sempre te la mateixa llargada, o bé si es tracta de números.

2.1 *Wrapper* genèric

El conjunt de regles necessàries per cada pàgina s'emmagatzemaran a la base de dades. Tindrem un *wrapper* genèric que a partir de les rutes i expressions regulars de cada element, podrà extreure la informació.

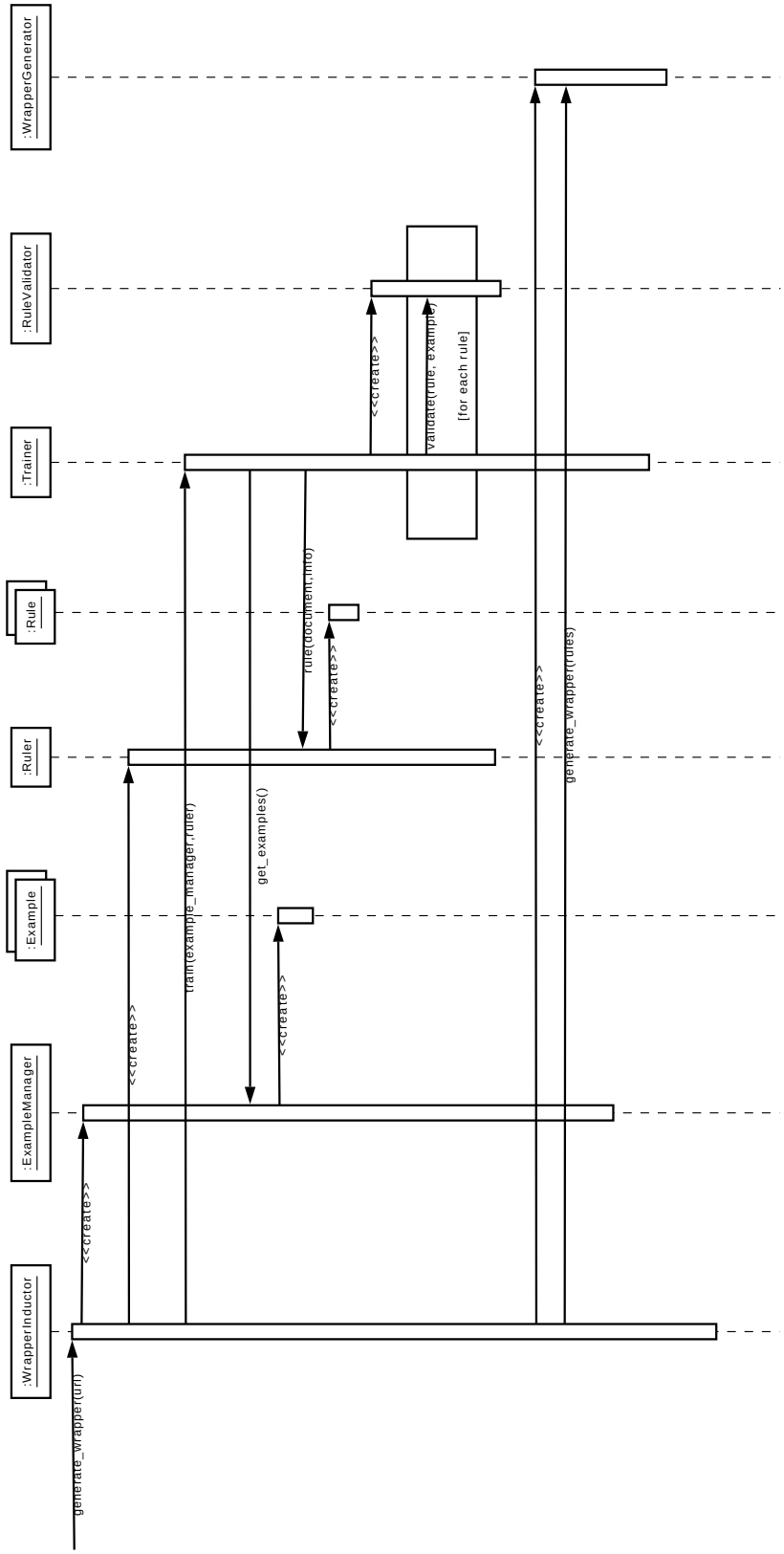

Figura 1: Diagrama de seqüència