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. Utilitzarem un dels exemples per generar el wrapper i en comprovarem la correctesa amb la resta. Tindrem en compte:

- 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.

  Haurem de tenir en compte, també, l'ordre de l'element respecte els seus *germans*.

  Utilitzarem tant etiquetes com atributs. Els principals atributs que mirarem seran `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 expresssions regulars.
  Podem crear aquestes expressions regulars mirant la resta de contingut de l'etiqueta. Es pot comparar amb la resta d'exemples disponibles per decidir quina és la informació rellevant.

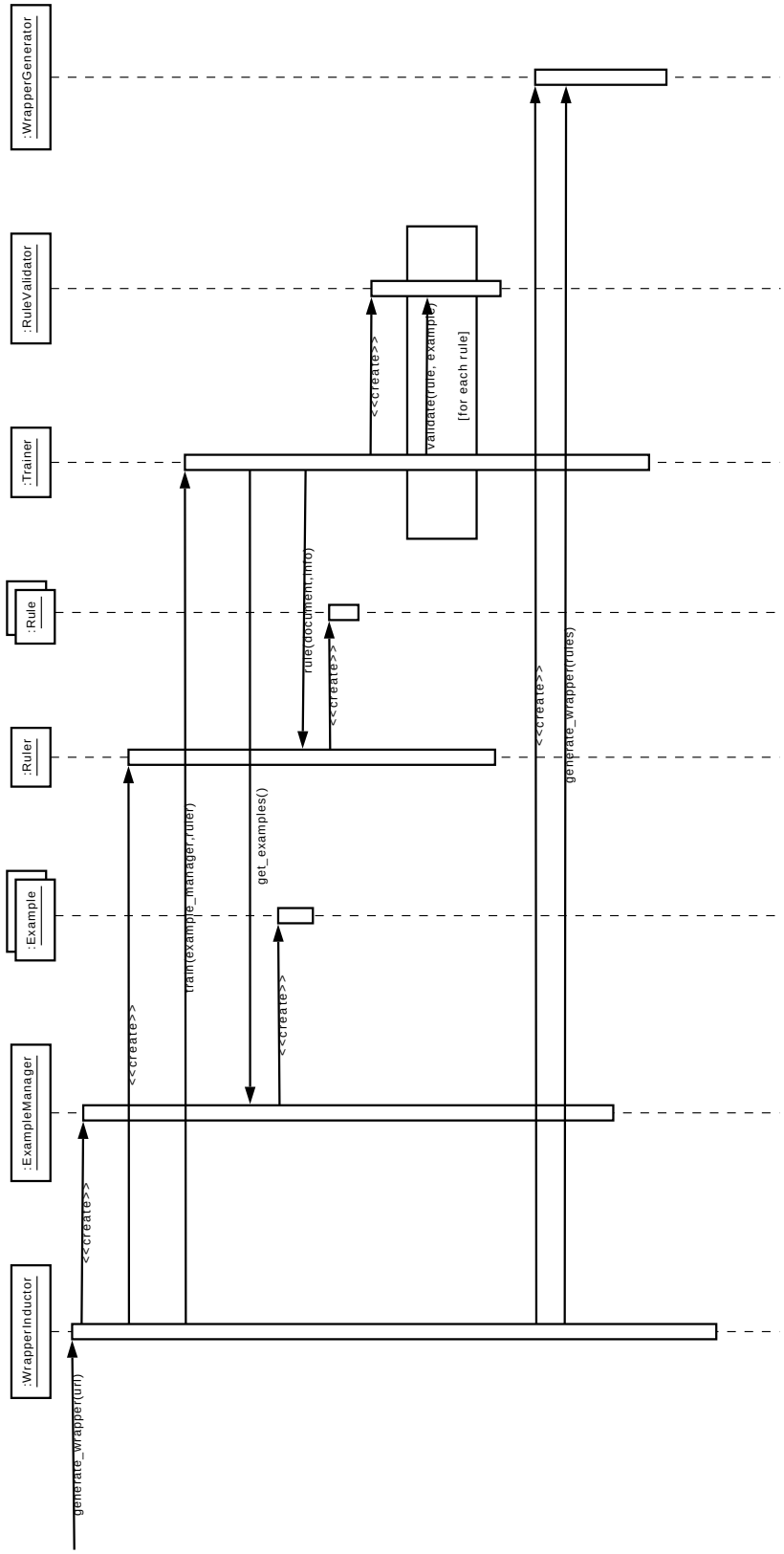

Figura 1: Diagrama de seqüència