Inférence de la structure d'une page web en vue d'améliorer son accessibilité

Encadré par : Y. Bonavero, M. Huchard et M. Meynard

Franck PETITDEMANGE

LIRMM
Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier

cnrs
dépasser les frontières

26 juin 2014

Sommaire

Sommaire

Accessibilité du web

Un enjeux sociétale important

Definition

Accessibilité Capacité d'accéder aux informations contenues dans une page et d'interagir avec.

Problèmes d'accessibilité (spécifique aux basses visions)

- Surcharge visuelle
- Police de caractère
- Contraste de couleur

Accessibilité du web

Besoin de comprendre la structuration d'une page

Problèmes des outils accessibilité

- Pas de traitement des couleurs locales
- Pas de prise en compte des profils utilisateurs

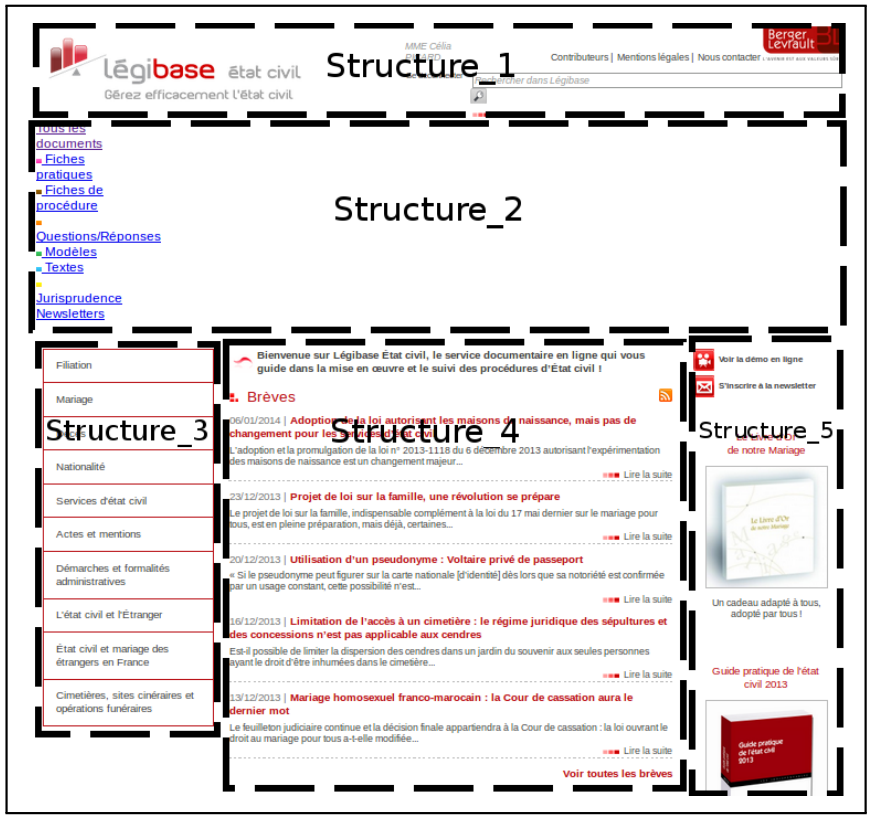
Besoin

- Comprendre les informations structurant une page web

Page web

Page web

- Technologies : HTML/CSS/Javascript
- Contenu hétérogène décrit par différentes structures logiques



Comment inférer les différentes structures logiques dans une page web ?

Difficultés

- Manque d'expressivité de HTML 4
- Pas de construction standard des structures logiques
- Écart entre la structure DOM et l'affichage dans un navigateur

Approche

- Étude des langages de publication de page web
- Étude des techniques d'extraction de structure d'une page

Sommaire

Évolution de la sémantique (1/2)

```
<ul class='menu'>
 <li><a href=".">l1</li>
 <li><a href=".">l2</li>
</ul>
<p class='menu'>
 <a href=".">l1</a>
 <a href=".">li2</a>
</p>
```

HTML 4

- Peu de sémantique
- Structure générique (DIV)
- Structure logique implicite

Évolution de la sémantique (2/2)

HTML 5

- Structure logique explicitée
- Sémantique pour décrire l'interface de la page est limitée

ARIA

- Ontologie d'une interface graphique
- Trop élaborée pour nos besoins mais est plus expressif

Évolution de la sémantique (2/2)

HTML 5

- Structure logique explicitée
- Sémantique pour décrire l'interface de la page est limitée

ARIA

- Ontologie d'une interface graphique
- Trop élaborée pour nos besoins mais est plus expressif

CSS
Un langage de mise en forme

Propriétés de mise en forme :

- avant-plan/arrière-plan
- police
- ...

Mécanisme de positionnement

- relatif
- absolu
- flottant

Synthèse

HTML 4 langage actuellement le plus exploité. Les inconvénients sont :

- la diversité de représentation d'une même structure logique
- la faible expressivité au regard des concepts décrits dans les pages web

Notre approche

- Proposer un Méta-modèle concrétisant mieux les concepts des pages et permettant de s'abstraire de la diversité de représentation des structures

Mapping

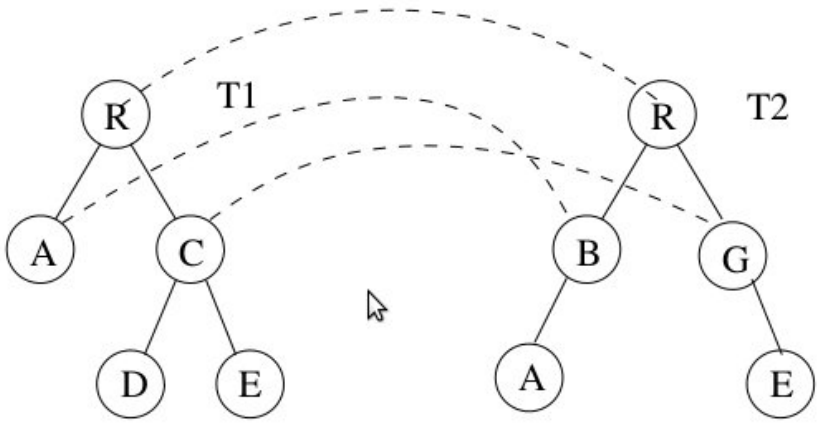
(Vieira et al., A fast and robust method for web page template detection and removal)

Mapping descendant restrictif

Permet de faire correspondre les plus grandes sous-structures communes entre deux arbres

Idée

Identifier les structures logiques des pages web par correspondance



Segmentation

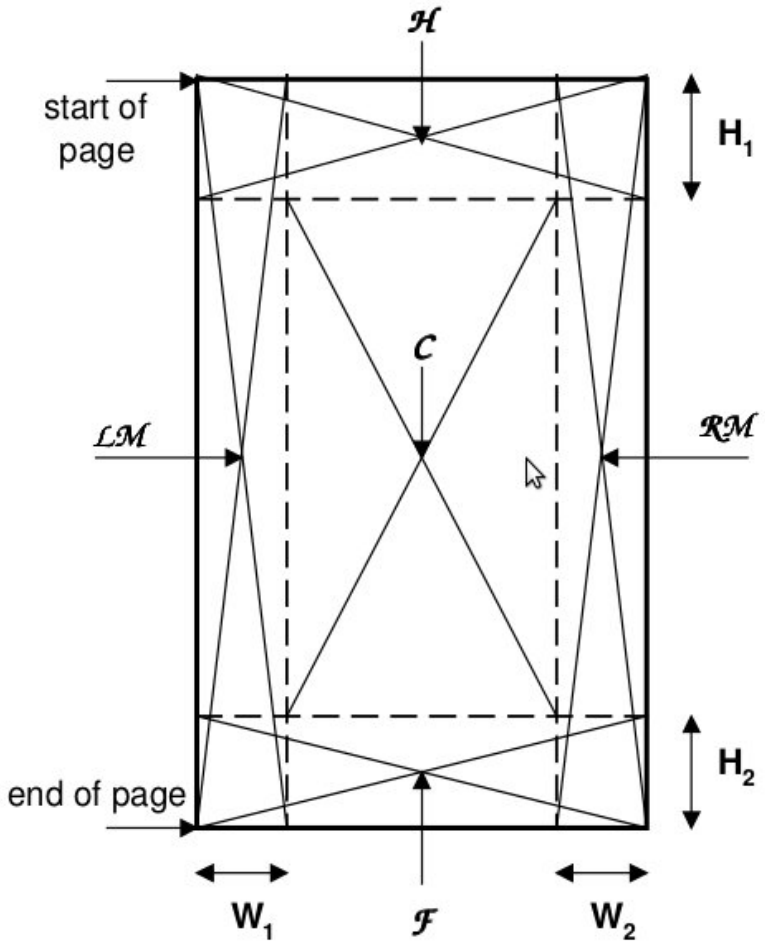
par pattern de présentation *(Milos Kovacevic et al., Recognition of Common Areas in a Web Page Using Visual Information : a possible application in a page classification)*

Observation

Les concepteurs de page web suivent approximativement les mêmes schémas de présentation

Idée

Regrouper les nœuds du DOM de la page suivant leurs coordonnées après la mise en page par le navigateur



Segmentation

par densitométrie textuelle *(Kohlschütter et al, A densitometric approach to web page segmentation)*

Étape 1 : identification de segments de petites tailles

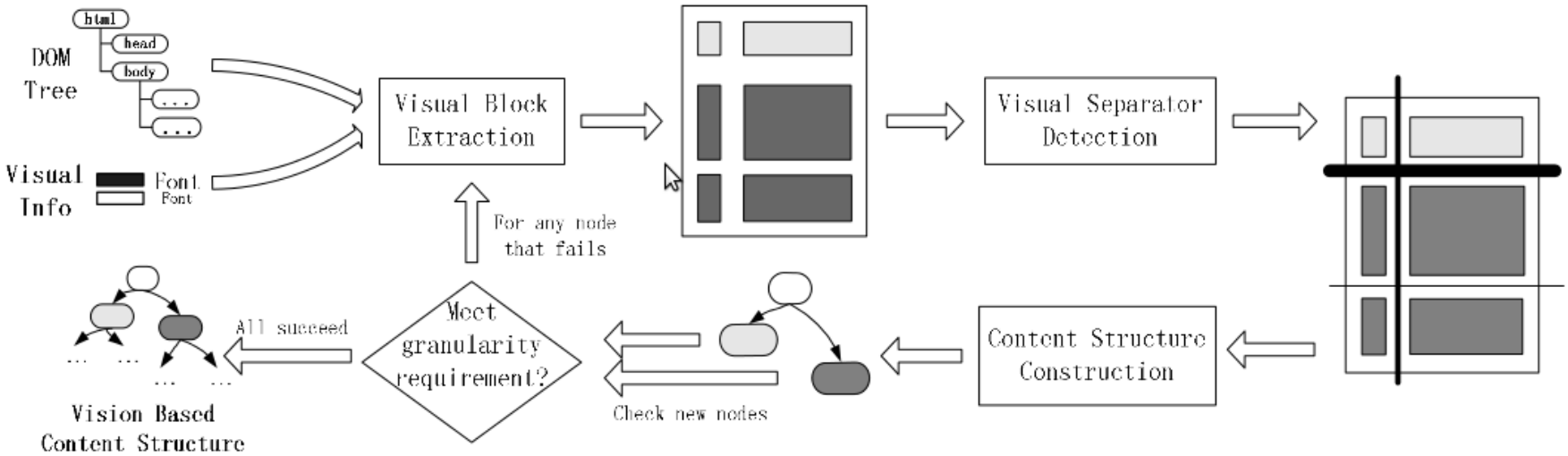
La page est vue comme une séquences de caractères entrelacés identifiés par des balises HTML. Les segments sont calculés d'après les variations dans le rythme des séquences.
Exemple :
Ici deux segments seront calculés
*<a>lienlien<a>lien un paragraphe *

Étape 2 : grossissement successif des segments par fusion

Les segments contingents dont la densité textuelle est proche sont fusionnées successivement.

Segmentation

par indice visuel *(Cai et al., Extracting content structure for web pages based on visual representation)*



Synthèse

Les inconvénients

- Mapping ne permet pas d'extraire la structure globale de la page
- La segmentation par pattern est trop dépendante de la présentation de la page
- La segmentation par densitométrie ne prend pas en compte les écarts possible entre le DOM et le rendu final
- Le calcule des séparateurs dans l'approche par indice visuel est une opération coûteuse $O(n^2)$

Notre approche : approche par segmentation par visuel

Propose un **découpage globale** de la page, **indépendant des patterns de présentation** et permet un **découpage fin** dans la structure d'une page.

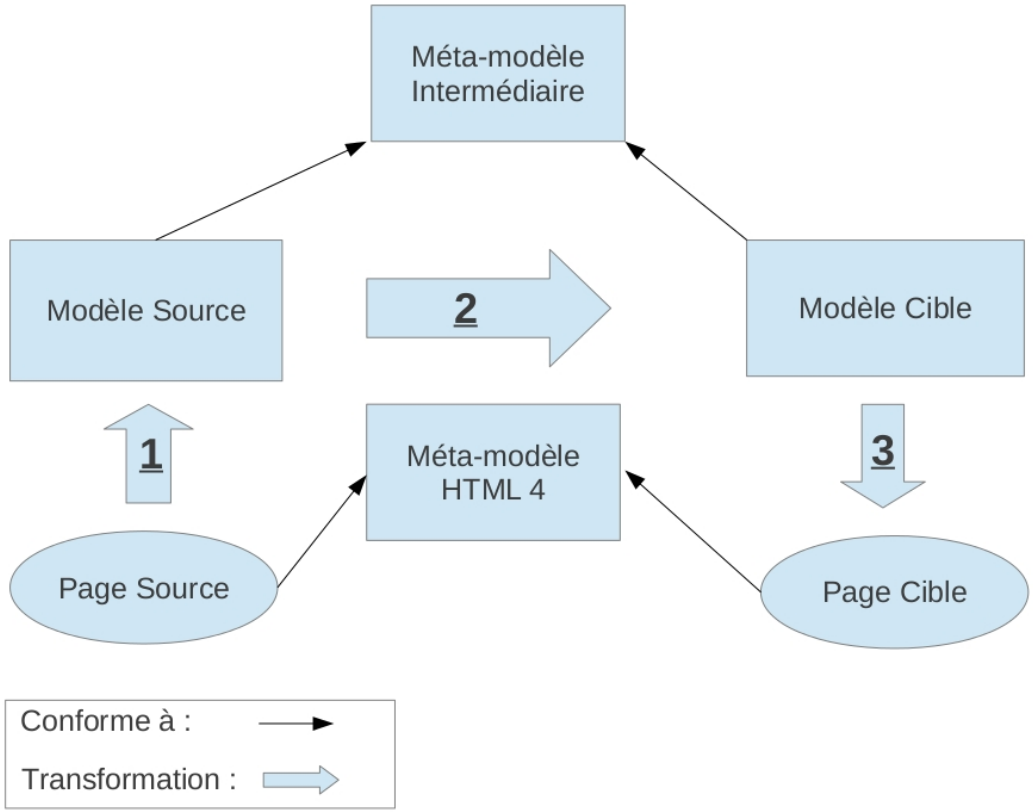
Sommaire

Approche générale

Une approche Ingénierie Dirigée par les Modèles

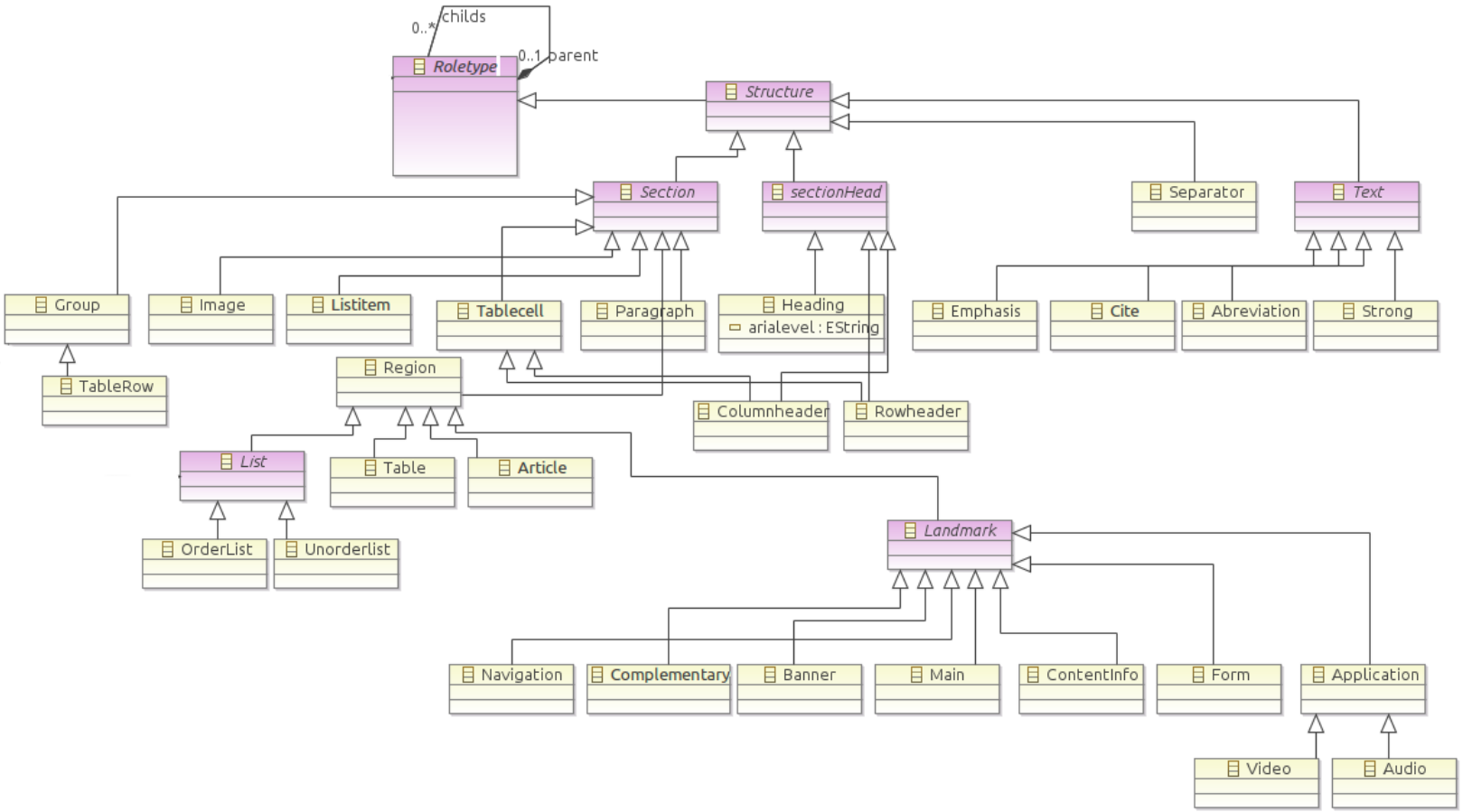
Avantages

- Meilleur expression des préférences
- Indépendant de la diversité de représentation des informations



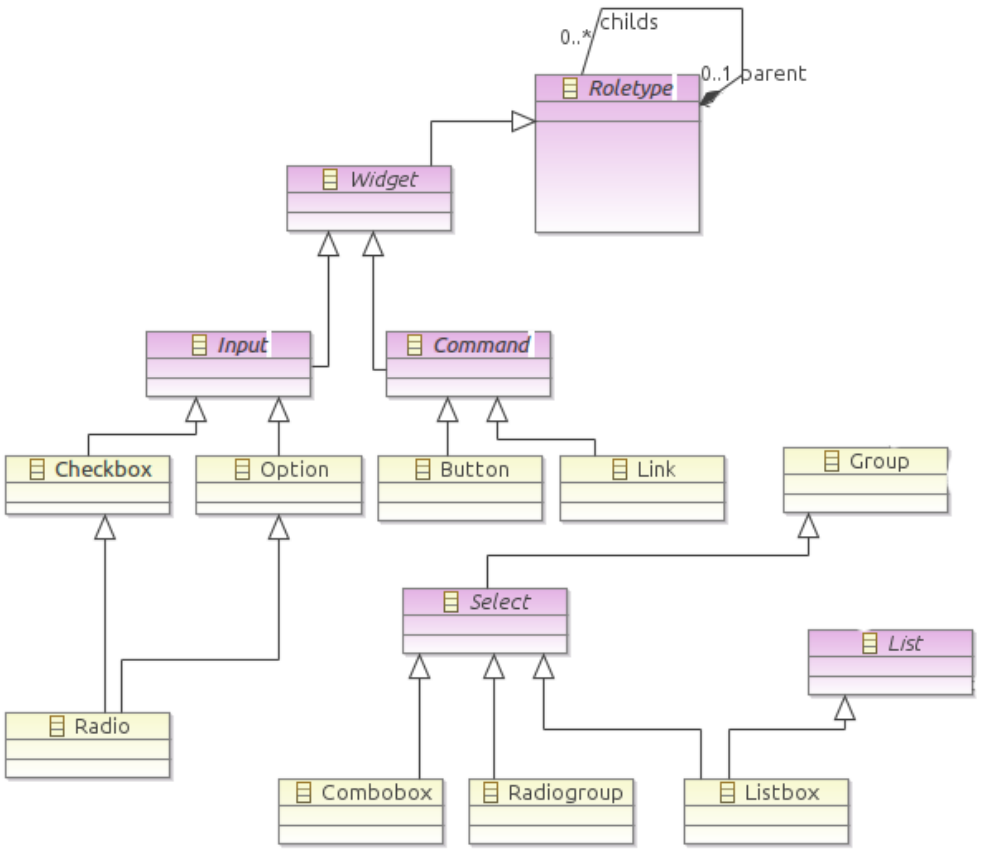
Méta-modèle intermédiaire

Éléments structurels



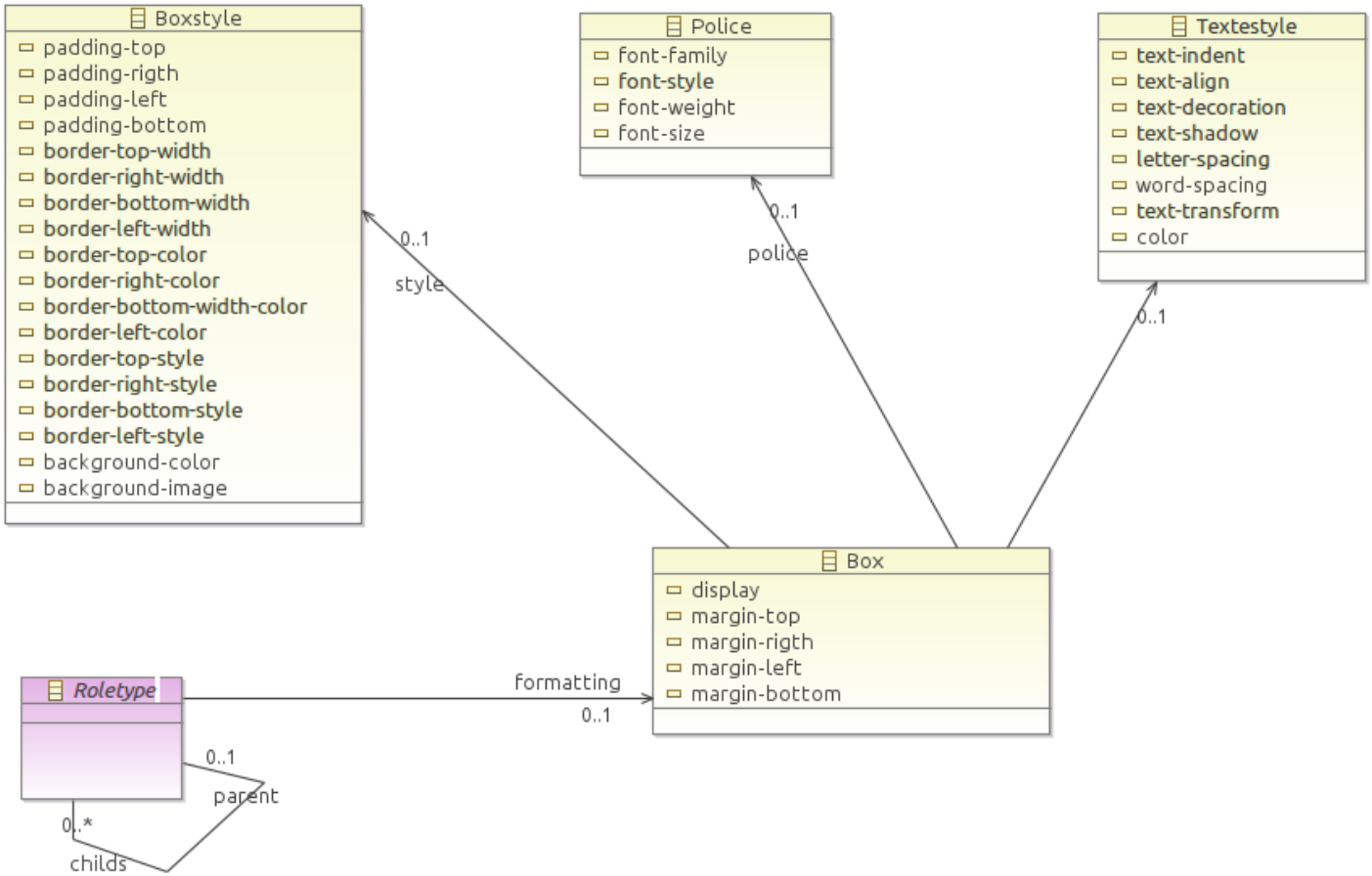
Méta-modèle intermédiaire

Éléments d'interaction



Méta-modèle intermédiaire

Éléments de mise en forme



Démarche générale

Concepts et heuristiques

Approche par fonctionnalité

Modèle orientée fonctionnalité

Approche par fonctionnalité

fonctions de détection proposées

Processus d'extraction

Segmentation globale page Berger-Levrault

Processus d'extraction

Segmentation locale page Berger-Levrault

Sommaire

Résultats

Résultats

- Proposition d'une approche IDM
- État de l'art sur les langages de publication de page web
- État de l'art des techniques d'extraction de structure
- Proposition d'un méta-modèle
- Adaptation et implémentation d'une méthode pour extraire les structures d'une page
- Proposition de pistes pour annoter les structures extraites

Difficultés et perspectives

Difficultés

- Domaine de recherche éloigné des connaissances de l'équipe de recherche
- Définir la problématique par rapport à la question de recherche (recherche de motifs d'intérêt dans un arbre DOM)
- Recherche d'articles traitant de la problématique

Perspectives

- Évaluation de la méthode d'extraction de structure
- Implémentation, évaluation et élargissement du processus d'annotation