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

26 juin 2014

Sommaire

Sommaire

Accessibilité du web
Un enjeu sociétal important

Définition

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 caractères
- Contraste de couleur

Accessibilité du web

Besoin de comprendre la structuration d'une page

Problèmes des outils d'accessibilité (spécifique aux basses visions)

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

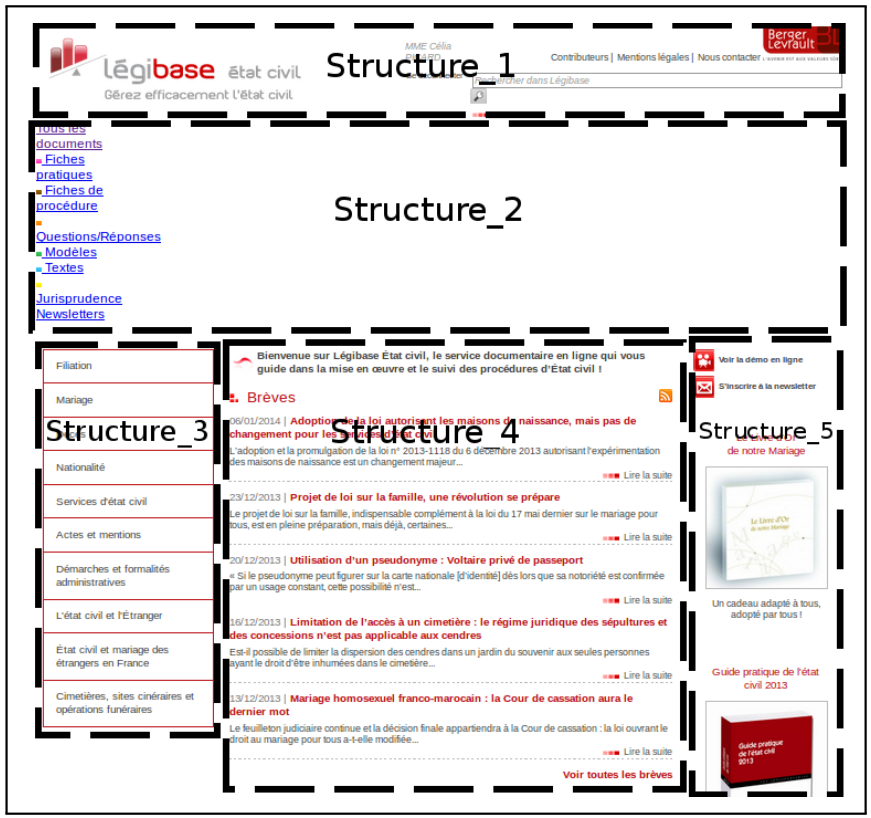
Besoin

- Comprendre les informations structurant une page web pour réaliser une adaptation basée sur cette structure

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=".">l2</a>
</p>
```

HTML 4

- Peu de sémantique
- Diversité de représentation
- Structure logique implicite

Évolution de la sémantique (2/2)

HTML 5

- Structure logique explicitée
- Sémantique pour décrire l'interface de la page 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 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
- l'absence de structuration explicite

Notre approche

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

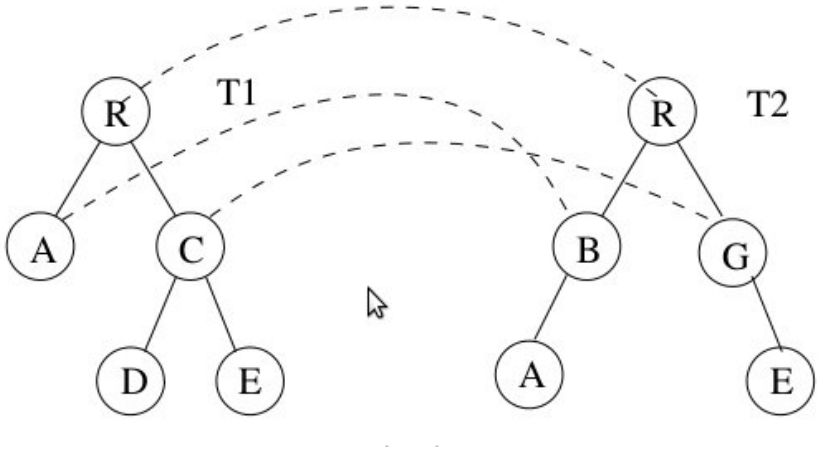
Mapping
(Vieira et al., A fast and robust method for web page template detection and removal)

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

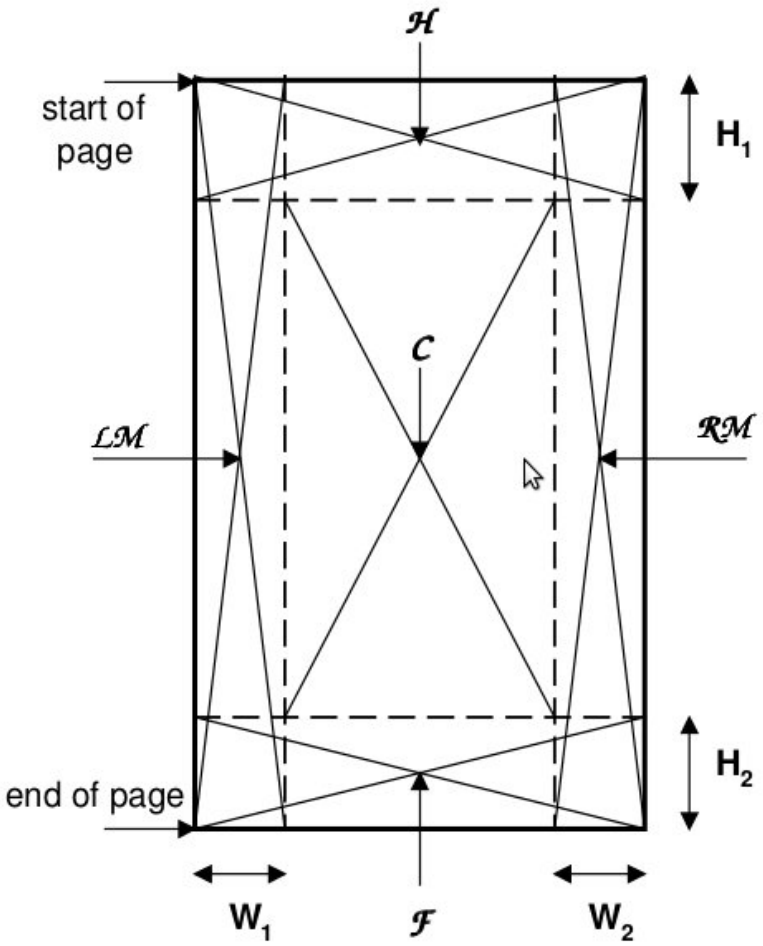
par pattern de présentation *(Milos Kovacevic et al., Recognition of Common Areas in a Web Page Using Visual Information : a possible application in a page classification)*

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

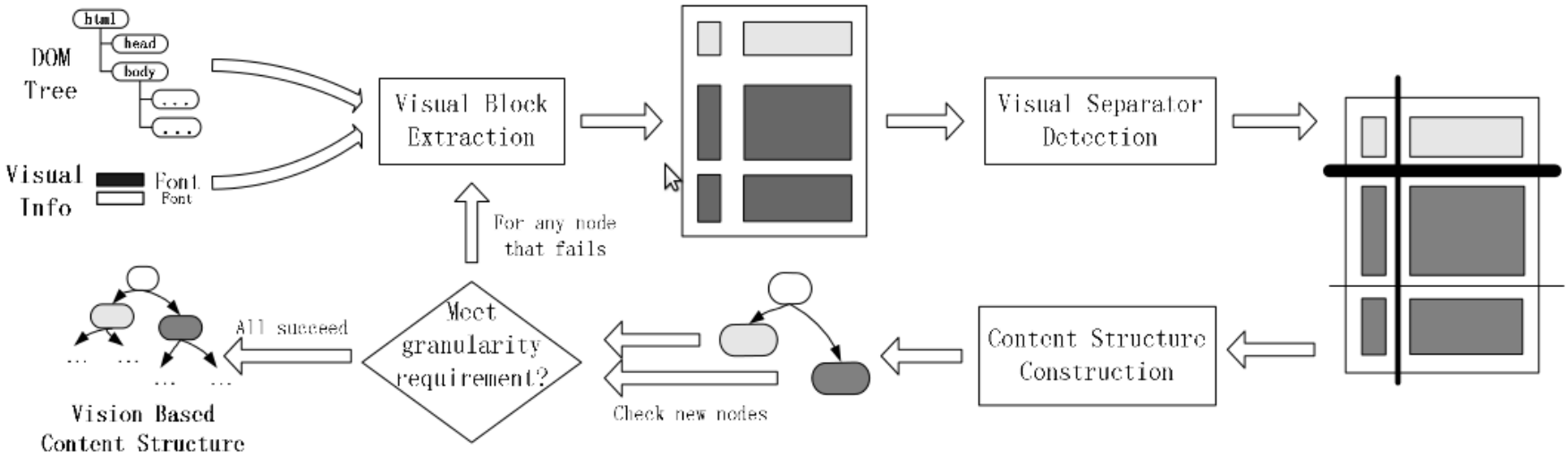
La page est vue comme une séquence de caractères entrelacés identifiés par des balises HTML. Les segments sont calculés d'après les variations dans le rythme des séquences pour lesquels on calcule une densité textuelle.
Exemple :
Ici deux segments seront calculés
*<a>lienAlienB<a>lienC un paragraphe *

Étape 2 : grossissement successif des segments par fusion

Les segments contigus 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)*



Idée

Regroupement par indice visuel

Synthèse

Les inconvénients

- Le *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 possibles entre le DOM et le rendu final
- Le calcul des séparateurs dans l'approche par indice visuel est une opération coûteuse $O(n^2)$

Notre approche : segmentation visuelle

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

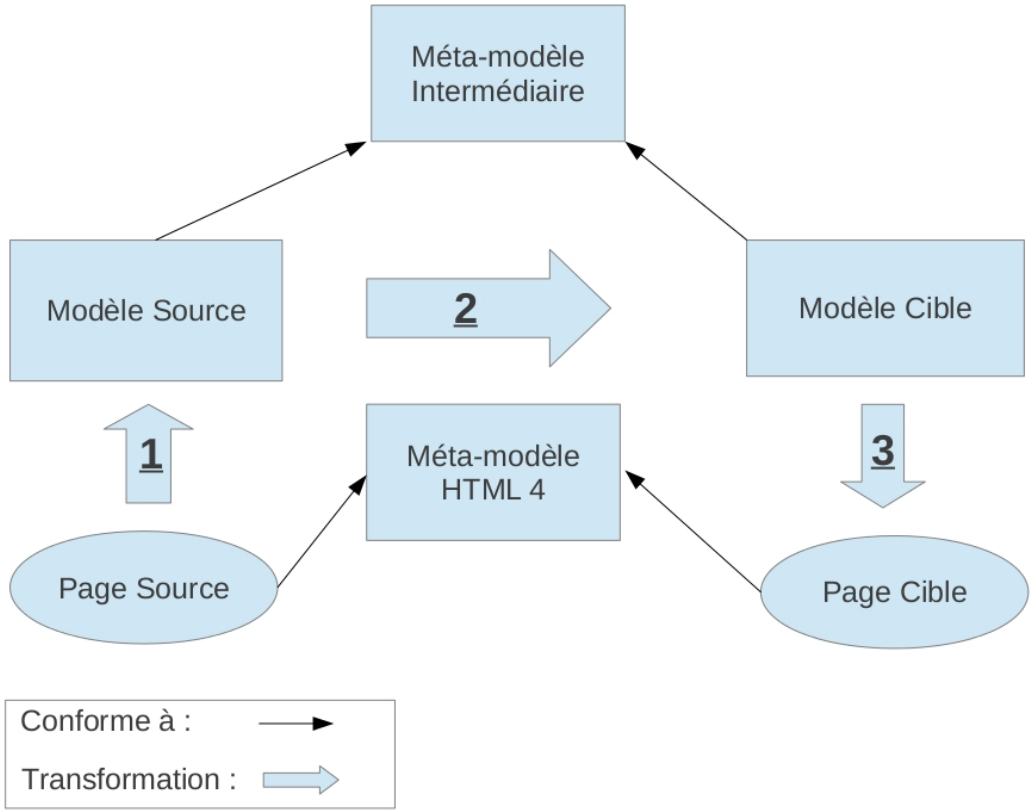
Sommaire

Approche générale

Une approche d'Ingénierie Dirigée par les Modèles

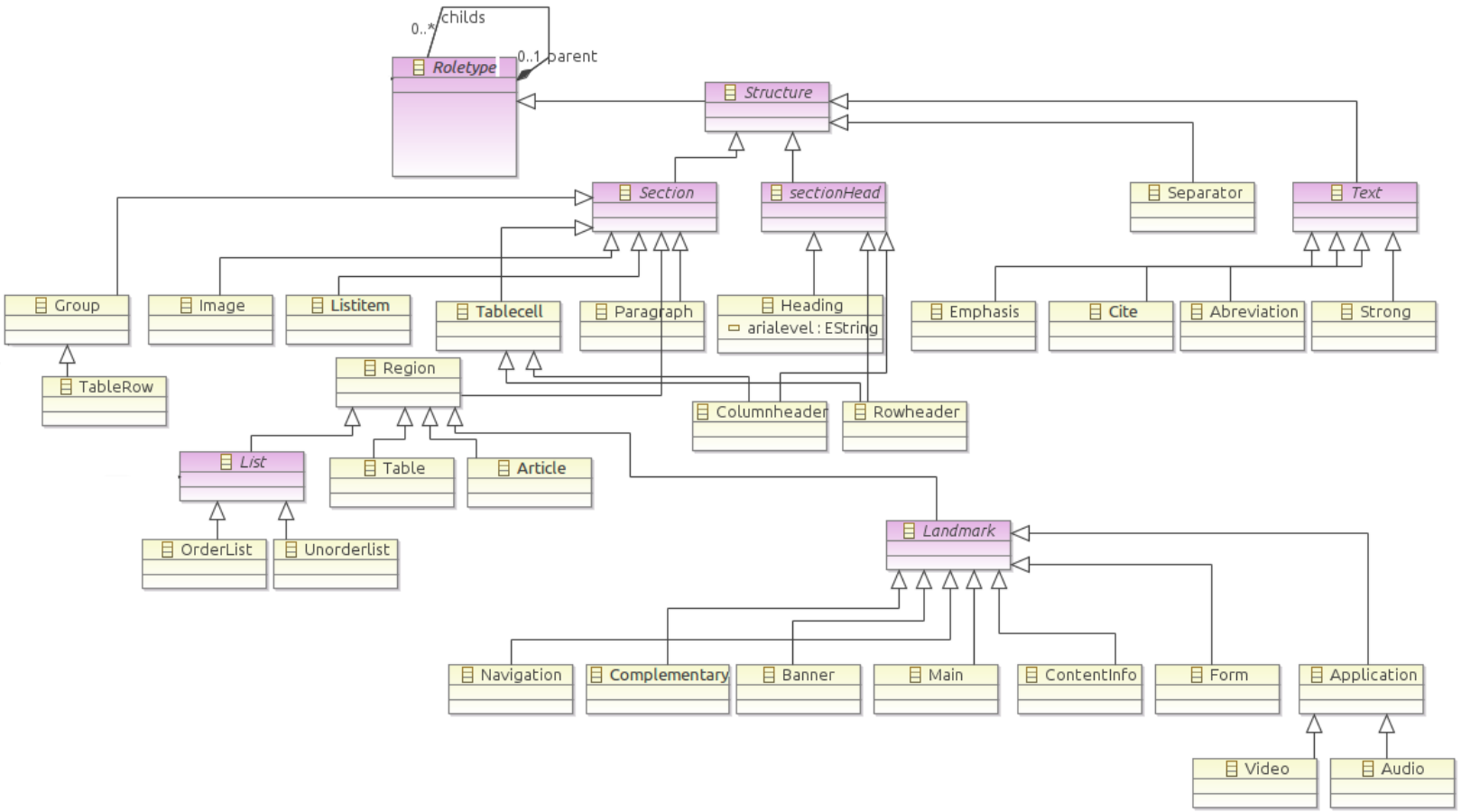
Avantages

- Meilleure expression des préférences
- Écriture d'adaptation indépendante des langages

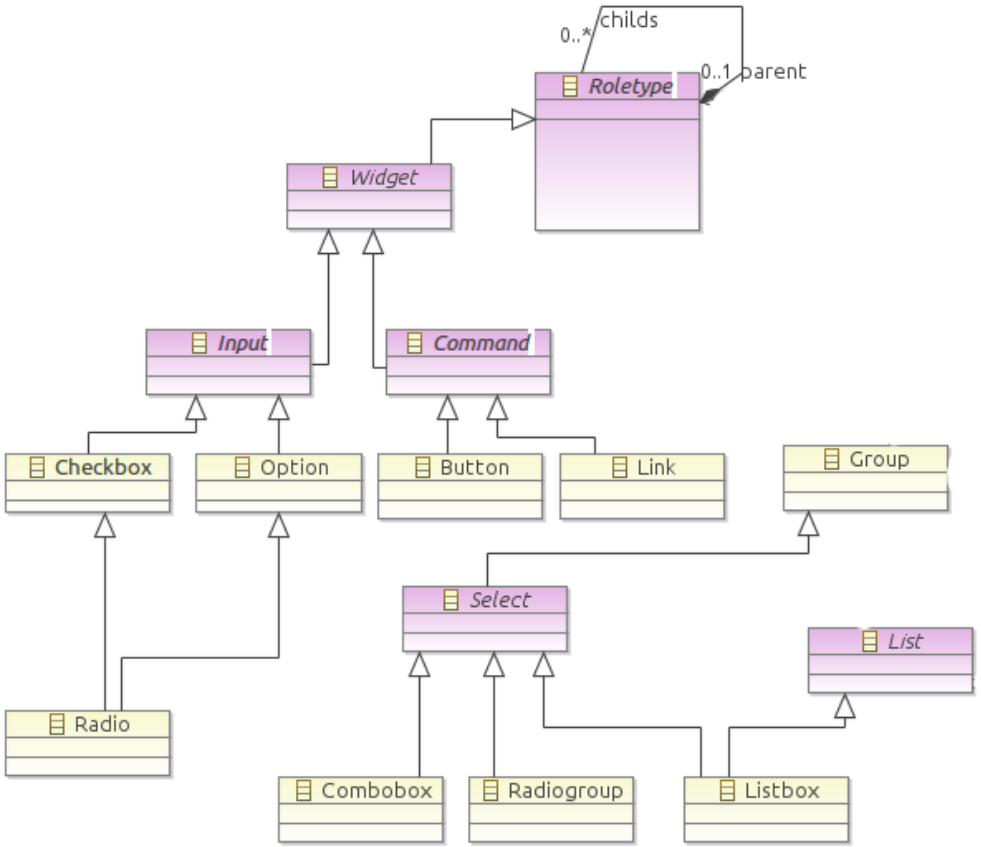


Méta-modèle intermédiaire

Éléments structurels

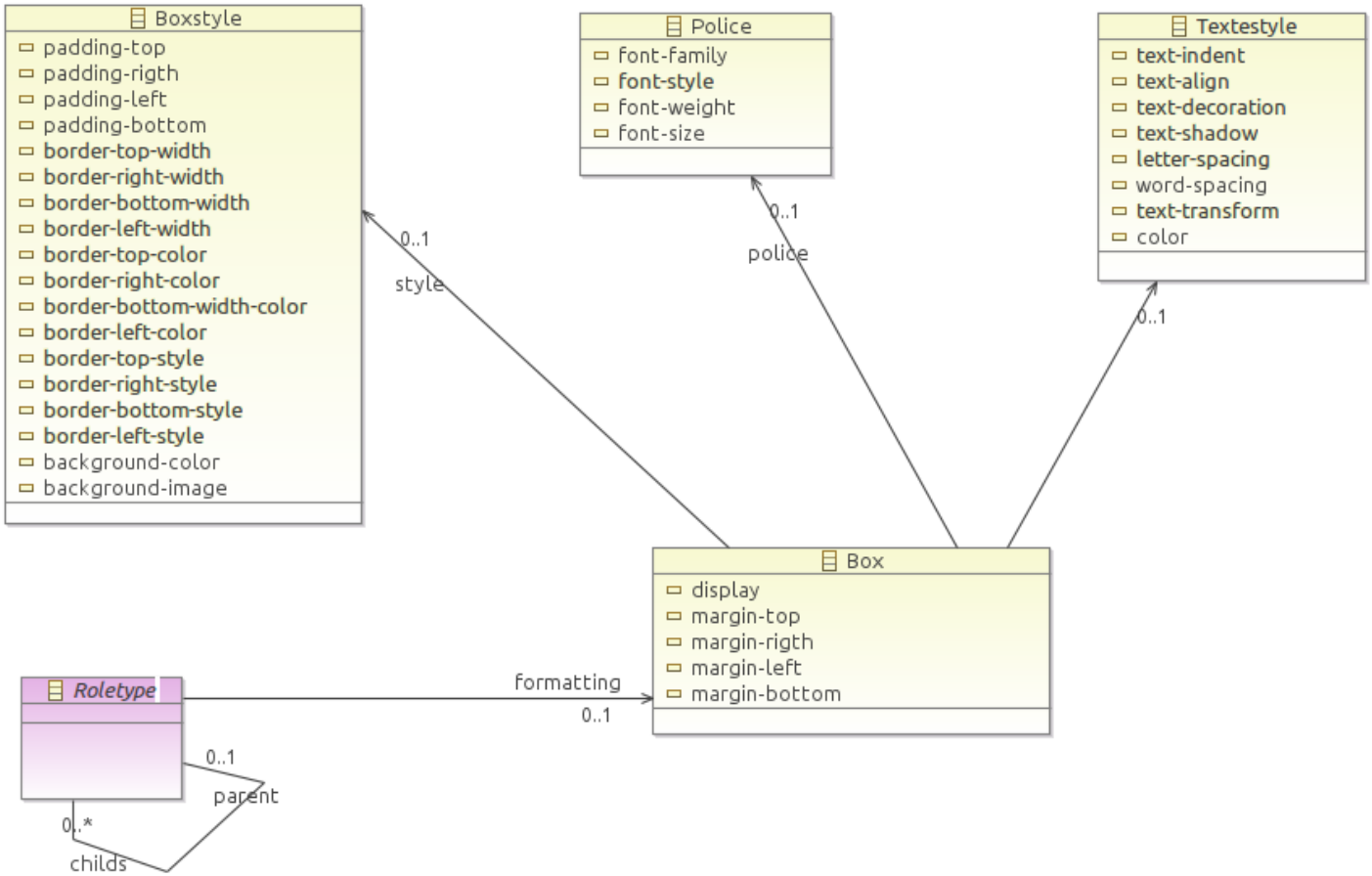


Méta-modèle intermédiaire
Éléments d'interaction



Méta-modèle intermédiaire

Éléments de mise en forme



Démarche générale

function ConstStructLog(*noeudDom*, *noeudStructInter*)
    poolNoeuds ← 0
DivisionDom(*noeudDom*, *poolNoeuds*)
    for i in *poolNoeuds* **do**
AppendChild(*arbreIntermediaire*, *poolNoeud*[*i*])
ConstStructLog(*poolNoeud*[*i*],*arbreIntermediaire*)
    end for
end function

function DivisionDom(*noeudDom*, *poolNoeuds*)
    if Divisible(*noeudDom*) == TRUE **then**
        for all Enfant de *noeudDom* **do**
DivisionDom(*noeudDom*, *poolNoeuds*)
        end for
    else
        poolNoeuds ← *poolNoeuds* + *noeudDom*
    end if
end function

Concepts et heuristiques

Concepts

Nœuds conteneurs : propriété CSS display égale à *bloc*

Nœuds de mise en forme : propriété CSS display à *inLine*

Taille d'un nœud : taille relatif à sa position dans le DOM et à la somme du poids des nœuds qu'il contient

Distance visuelle : signale les changements de propriété de mise en forme des nœuds

Règles

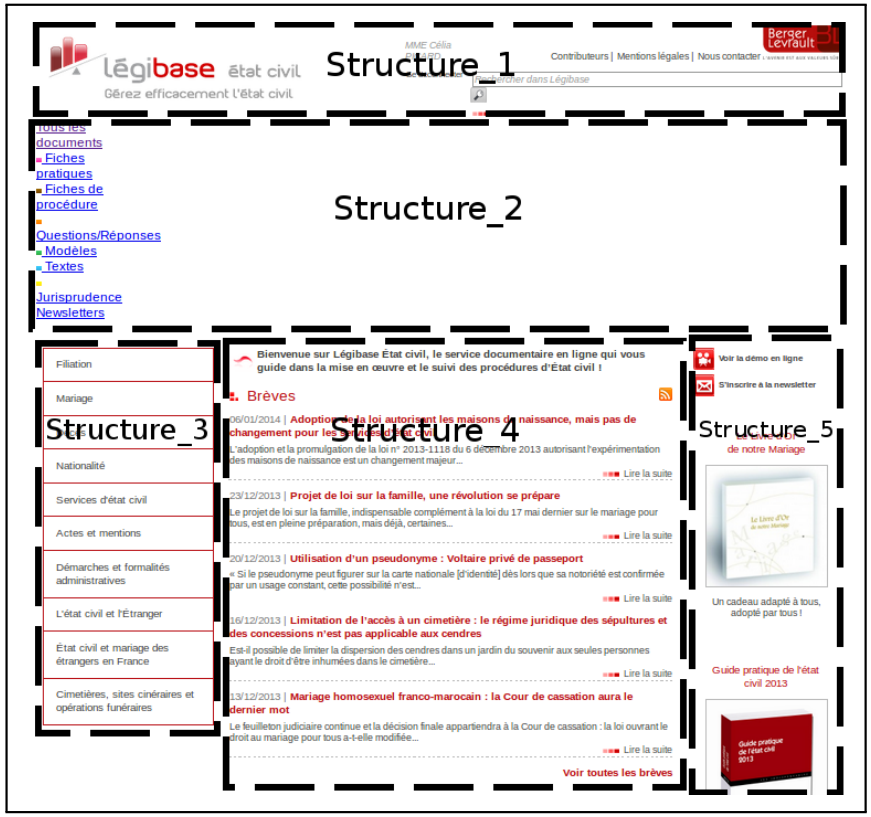
Règle 2 : si le nœud possède un seul enfant et que cet enfant n'est pas un nœud de données alors on parcourt ce nœud

Règle 4 : si l'un des enfants du nœud est un nœud conteneur et que sa taille est supérieure à un certain seuil alors on parcours ce nœud

Règle 8 : si la fonction de distance visuelle est vérifiée pour l'un des nœuds enfants alors on parcourt ce nœud et on extrait les enfants qui vérifient la fonction

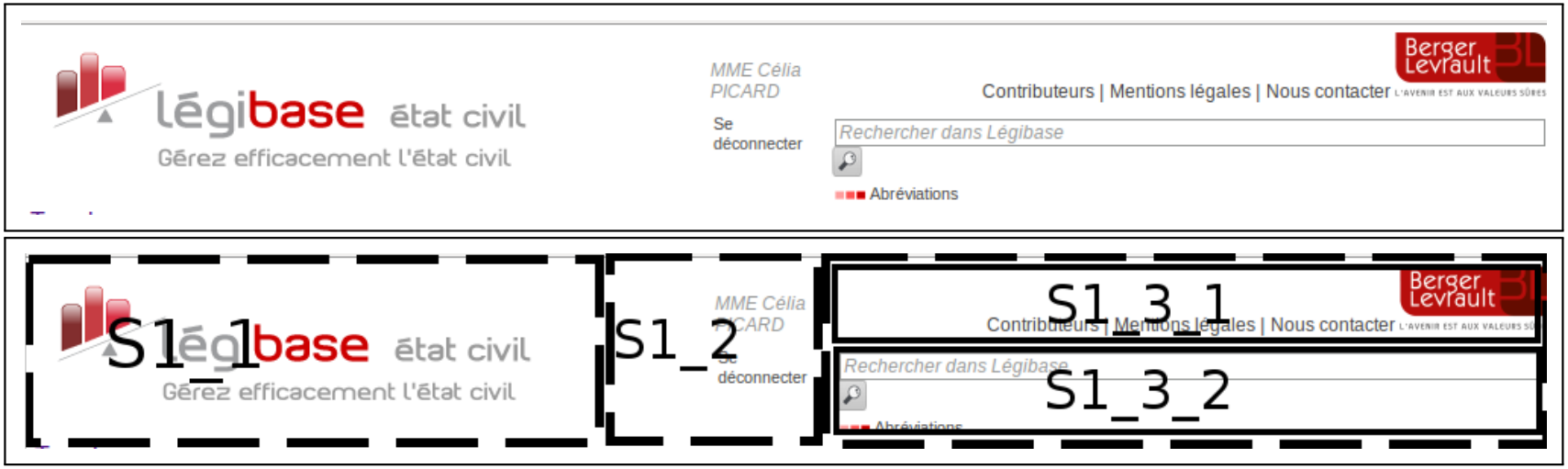
Processus d'extraction

Segmentation globale page Berger-Levrault



Processus d'extraction

Segmentation locale page Berger-Levrault



Approche par fonctionnalité

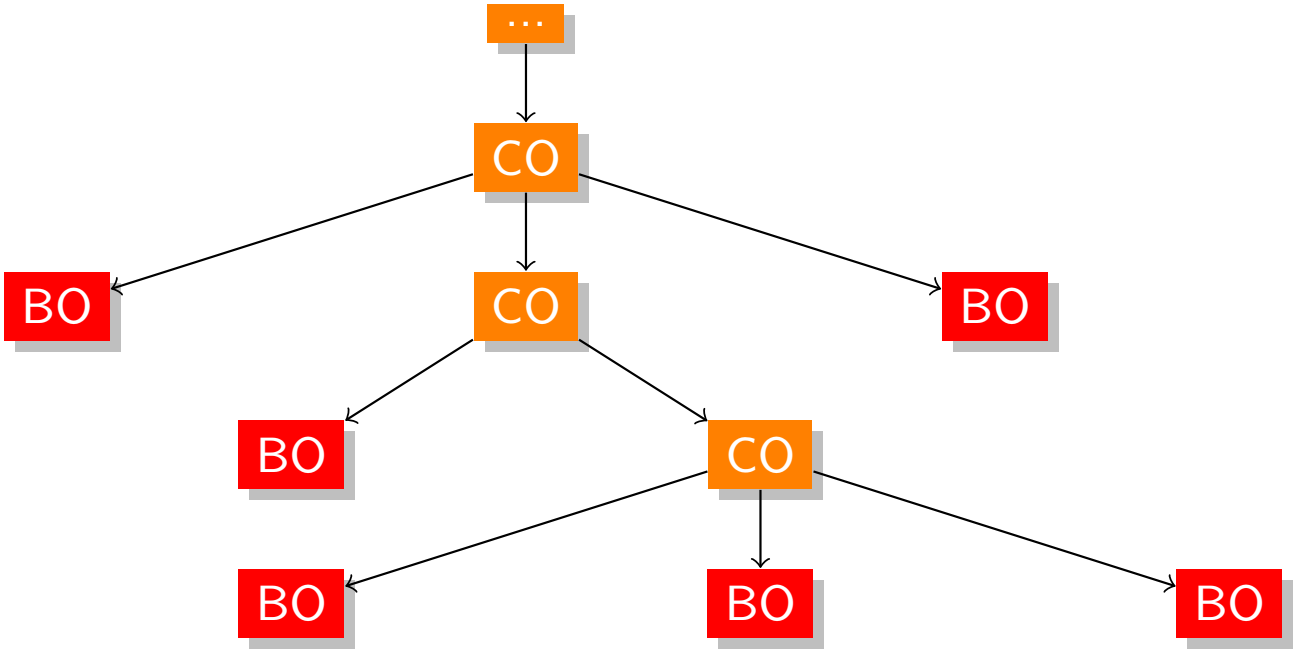
Construction de fonction d'annotation *(Chen et al., Function-based object model towards website adaptation)*

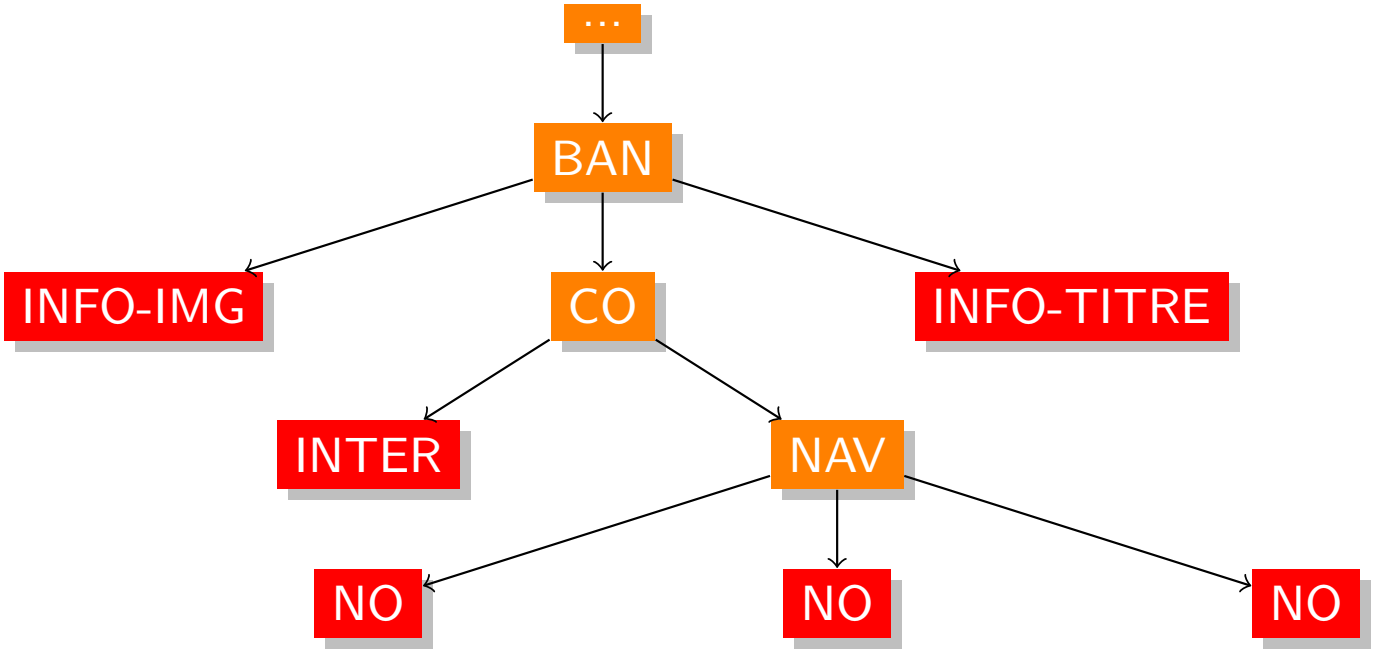
Classe de nœuds :

- Informatif
- Navigation
- Interaction
- Décoration

Chaque classe possède les propriétés :

- Hyperlien
- Sémantème
- Décoration
- **Position**


...
CO
BO
CO
BO
BO
CO
BO
BO
BO


...
BAN
INFO-IMG
CO
INFO-TITRE
INTER
NAV
NO
NO
NO

Sommaire

Difficultés

Difficultés

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

Résultats

Résultats

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

Perspectives

Perspectives à court terme

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

Perspectives à long terme

- Évaluation des modèles intermédiaires générés par le processus d'extraction et d'annotation sur un grand nombre de pages
- Acquisition des préférences utilisateurs
- Intégrer les préférences utilisateurs dans le processus de transformation