

## Abstract

We present an overview of a prototype we are developing for in the context of web archives (page comparison, crawling and information retrieval). It analyses pages based on their DOM tree information and their visual rendering. This tool implements a modified version of VIPS with the aim of enhancing the precision of visual block extraction and the hierarchy construction, from the user perspective. First, the visual rendering of a page, produced by several browsers, is segmented into rectangular blocks. Then, the extracted blocks are analysed looking for visual overlaps, which are analysed using a adapted version of the XY-Cut algorithm and computational gometry methods and resolve the overlap. As a result we may have different shapes of blocks, rectangular and non-rectangular blocks (polygons). Finally, the visual block tree, representing the layout of the page is analysed in order to have a more coherent layout disposition.

## The Problem

**MOST POPULAR**

1. Fanning forum, Justice Scalia says appeals court judge led...

2. Anti-Japan protests erupt across China on occupation anniversary...

3. Special Report: China's car makers out...

4. "Swing" again linked to risky sex among...

5. Romney denies Obama supporters in...

Ten years ago, no discerning Chinese consumer would have bought China-designed cars. Not only were they accused of being legal counterfeit of foreign models, but the quality and safety were also mistrusted.

Now, despite their homely looks, some indigenous models are striking a balance between no-frills affordability and acceptable quality. In China, it is the age of the good-enough car. And this has potentially significant implications for the world auto industry.

Models such as the Panda and the new Haval H3 are becoming popular not only in China but increasingly in emerging markets, from Indonesia to Egypt and Ukraine. They are driving China's auto exports to record levels, even as growth in China's auto market slows down.

GETTING TRACTION ABROAD

Exports of Chinese-produced vehicles are forecast by China's auto association to hit one million vehicles in 2012, up from 849,500 vehicles last year. Some automotive analysts are predicting a 50 percent increase to 1.25 million vehicles.

China's cars are very to China

China's cars are very to China

China's cars are very to China

Does this date belongs to block 2.5 or block 2.6?

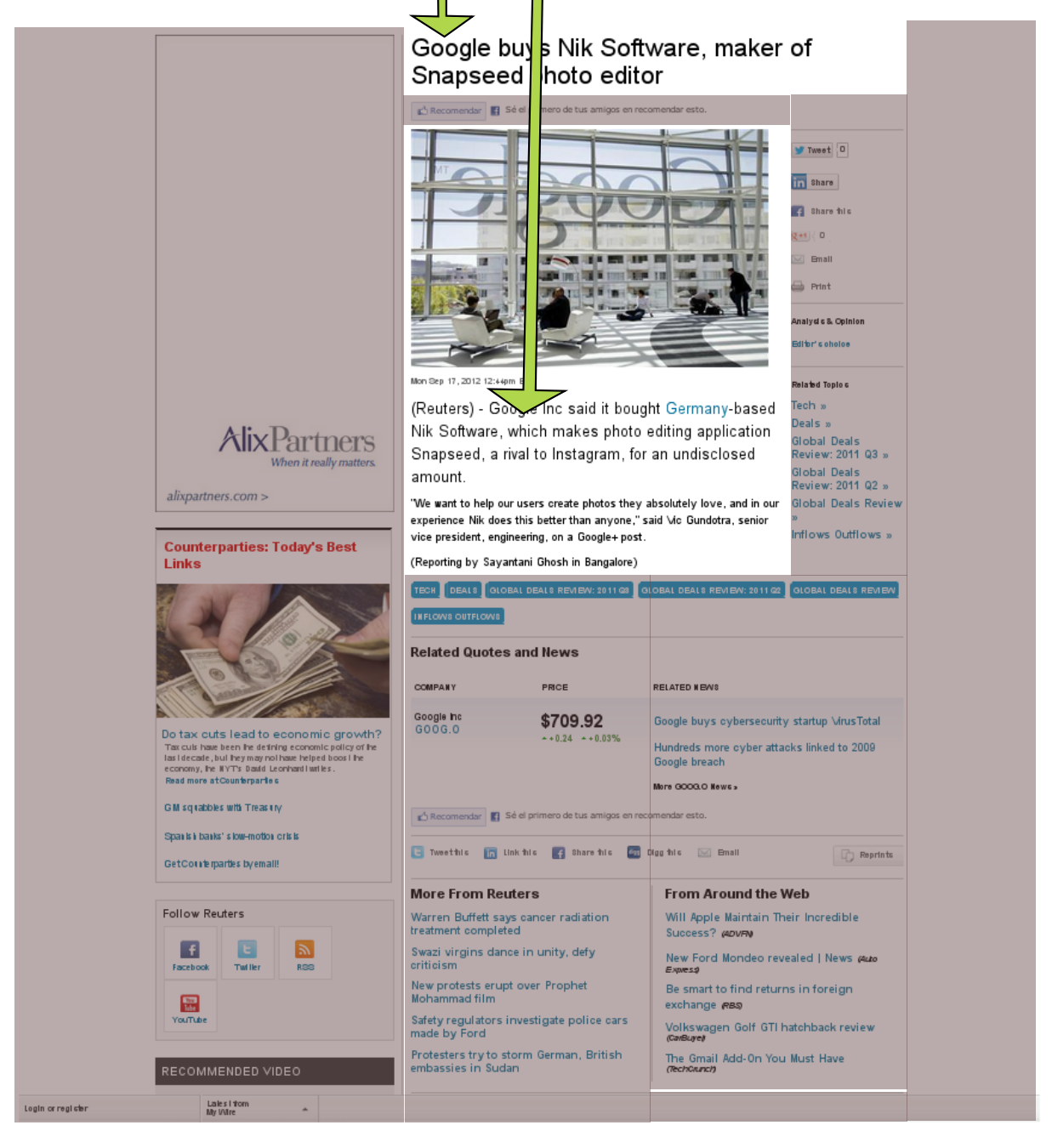
A rectangular shape for blocks is not always the most suitable, because overlapping may occur, making the description of a page ambiguous: « to which block should this date belongs ? »

## Why to Segment a Web Page?

Web Page segmentation is a technique used for dividing a web page into particular parts, not overlapping, called segments or blocks.

### Describing the page as a unit

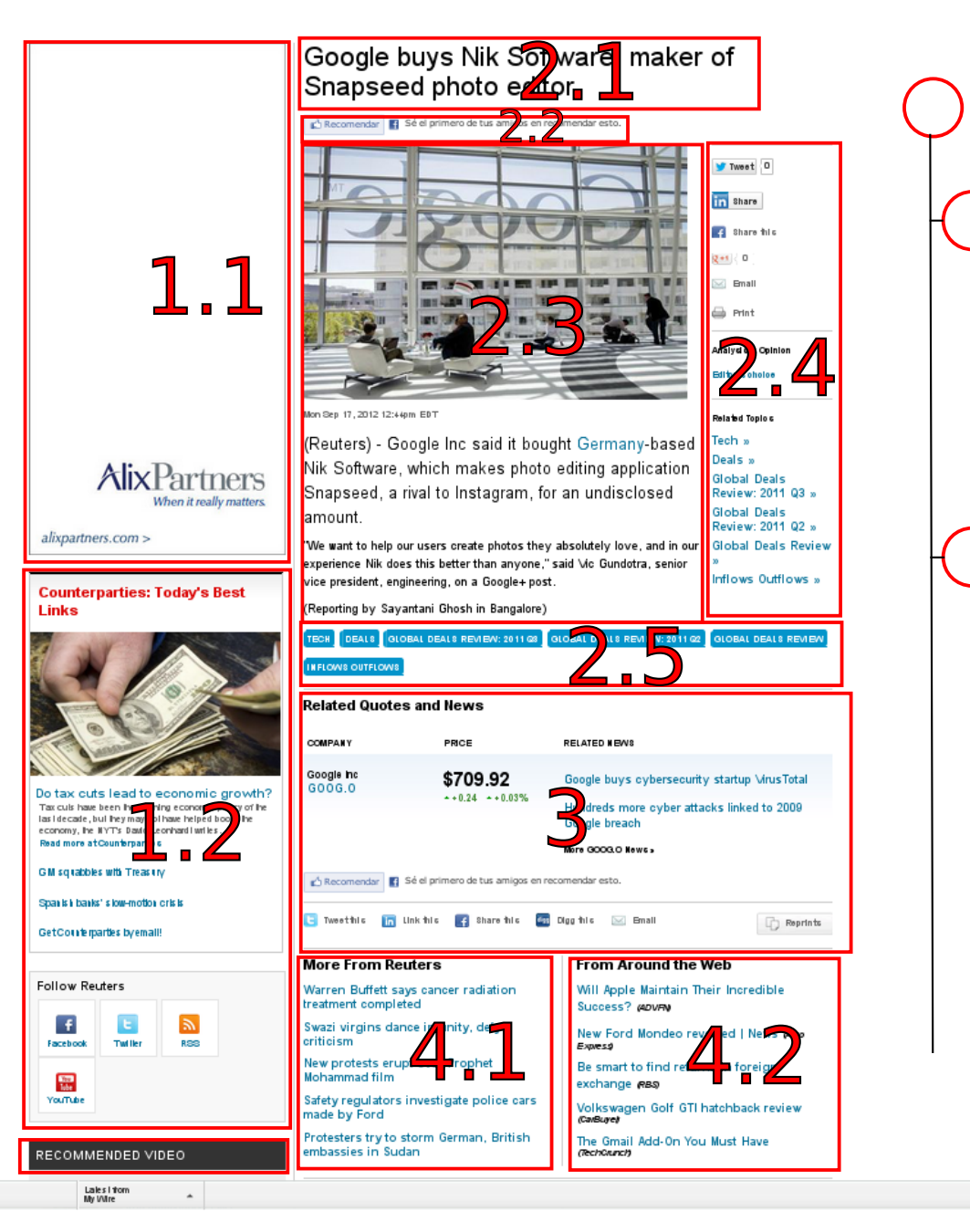
Relevant content



### Describe the page based on detected visual blocks

visual blocks hierarchy

- page
  - VB1
    - VB1.1
    - VB1.2
  - VB2
    - VB2.1
    - VB2.2
  - ...



- Describe the content of a web page as a whole can lead to a not so relevant result : content not related to main content are included in the description
- Noisy content is included in the analysis
- Dividing the page in segments or blocks allows to treat each one as a « subpage », most important blocks being representative of the whole page
- Noisy content can be excluded
- Blocks 2.1 and 2.2 are the most importants in the above example

## Our thesis

**MOST POPULAR**

1. Fanning forum, Justice Scalia says appeals court judge led...

2. Anti-Japan protests erupt across China on occupation anniversary...

3. Special Report: China's car makers out...

4. "Swing" again linked to risky sex among...

5. Romney denies Obama supporters in...

Ten years ago, no discerning Chinese consumer would have bought China-designed cars. Not only were they accused of being legal counterfeit of foreign models, but the quality and safety were also mistrusted.

Now, despite their homely looks, some indigenous models are striking a balance between no-frills affordability and acceptable quality. In China, it is the age of the good-enough car. And this has potentially significant implications for the world auto industry.

Models such as the Panda and the new Haval H3 are becoming popular not only in China but increasingly in emerging markets, from Indonesia to Egypt and Ukraine. They are driving China's auto exports to record levels, even as growth in China's auto market slows down.

GETTING TRACTION ABROAD

Exports of Chinese-produced vehicles are forecast by China's auto association to hit one million vehicles in 2012, up from 849,500 vehicles last year. Some automotive analysts are predicting a 50 percent increase to 1.25 million vehicles.

China's cars are very to China

China's cars are very to China

China's cars are very to China

- Using convex polygons to model blocks allows having a more precise segmentation, therefore a better description of the page content
- Classify blocks taking as reference layout objects taxonomy (header, footer, etc) allows to order the blocks in a hierarchy

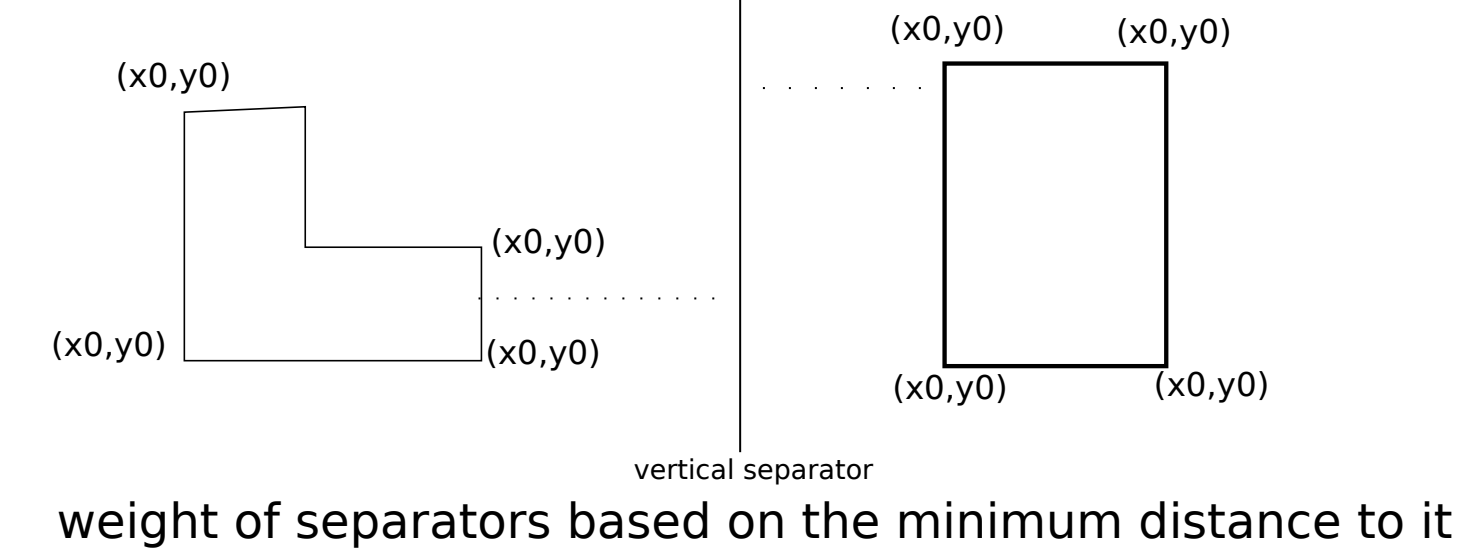
### Hybrid Approach Segmentation Algorithm

```
function: segment(node)
root_block = detect_blocks(node)
root_block = detect_separators(root)
root_block = detect_overlapping(root)
root_block.merge_separators
hierarchy = root_block.detect_layout
end

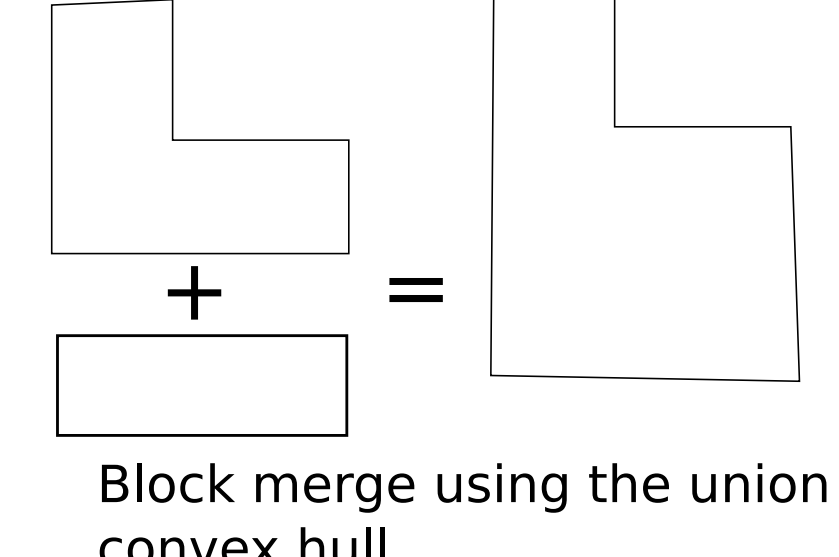
function: detect_overlapping(block)
for each pair of children of block (b1,b2)
if detect_overlapping(b1,b2)
window = define_window_area(b1,b2)
new_blocks = apply_image_segmentation(window)
for each block in new_blocks (nb)
if b1 contains nb, delete b1 and keep nb
if area(b1) - area(nb) > kA, delete nb and keep b1
if nb contains b1 and b2, keep nb and delete b1 and b2
in other case add nb to children list
//same procedure for b2
end
end
end
```

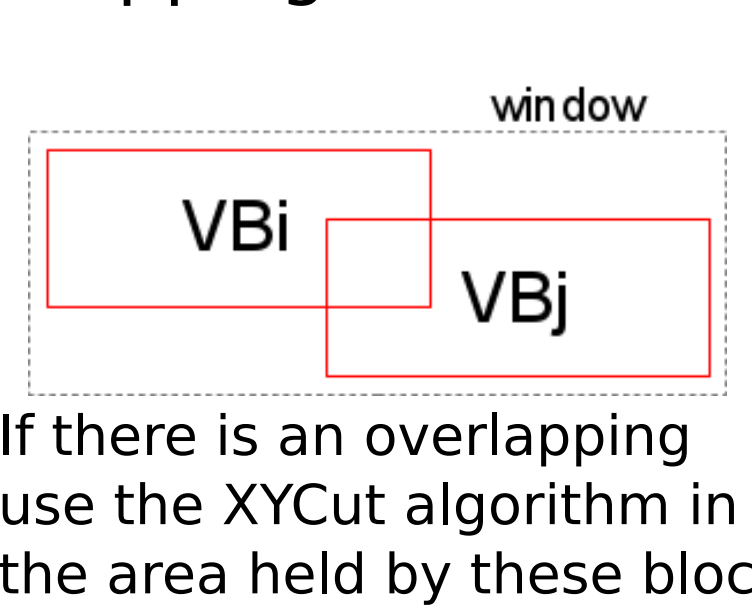
## About the Prototype

Our prototype aims to have a more precise web page segmentation using polygons to denote blocks. It is based on VIPS algorithm with some extra features:

- Polygon based model for visual blocks and separators description:
$$\text{block} = \{(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)\}$$
$$\text{page} = \{\text{block}_1, \dots, \text{block}_n\}$$
$$\text{separator} = \{(x_{sp}, y_{sp}), (x_{ep}, y_{ep})\}$$
- Computational geometric methods for visual blocks and separator detection:

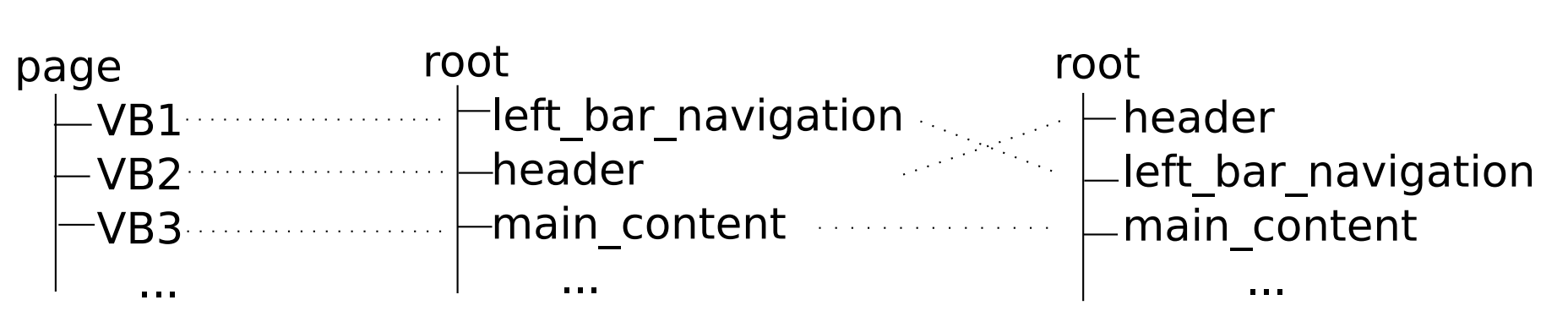
weight of separators based on the minimum distance to it



Block merge using the union convex hull
- Overlapping detection and resolution:

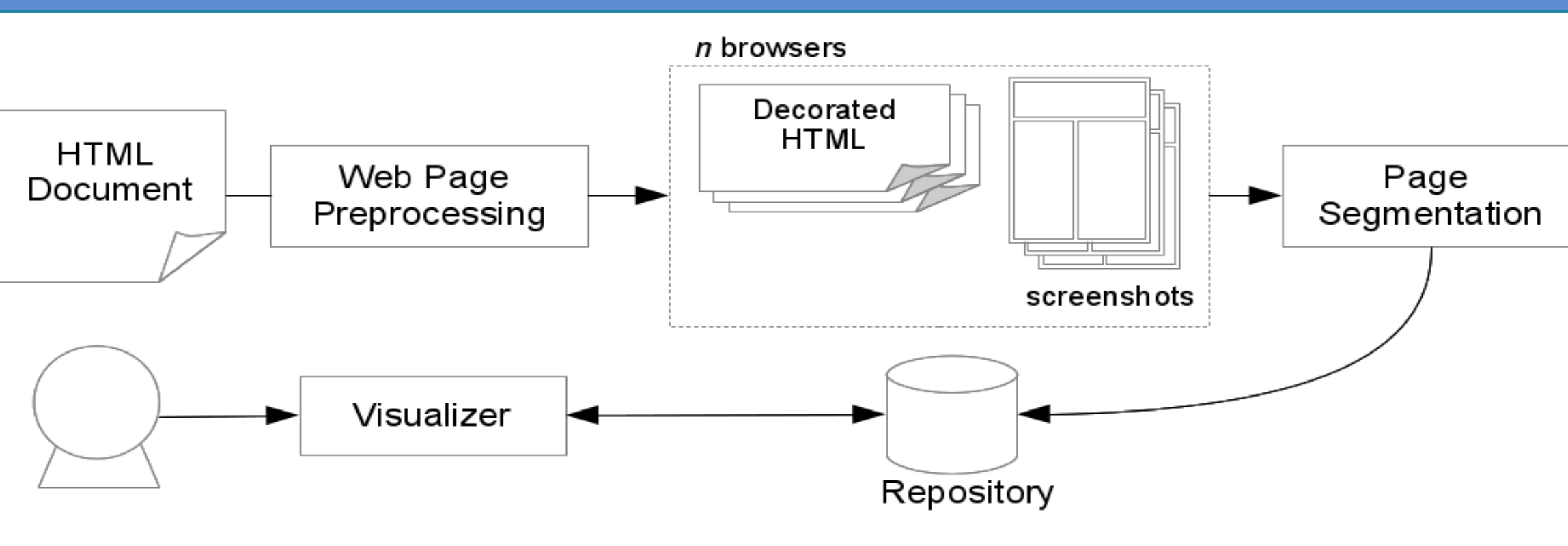
If there is an overlapping use the XYCut algorithm in the area held by these blocks

then...

  - If detected polygon is contained in original block, delete block and replace it by polygon
  - If polygon shape is similar to the original block, and the geometry not much different than a threshold, keep original
  - else keep original
  - if polygon contains more than one original block, the latter are deleted
- Page layout comparison with predefined templates for determine blocks order in a hierarchy:

page root left\_bar\_navigation header left\_bar\_navigation main\_content

## General Process



**Input:**

- Decorated HTML  
It is the original source document without tags that give no information for the segmentation and for each remaining tags, extra attributes are added to denote the visual cues

Original HTML tag

```
<div id="item">
...
</div>
```

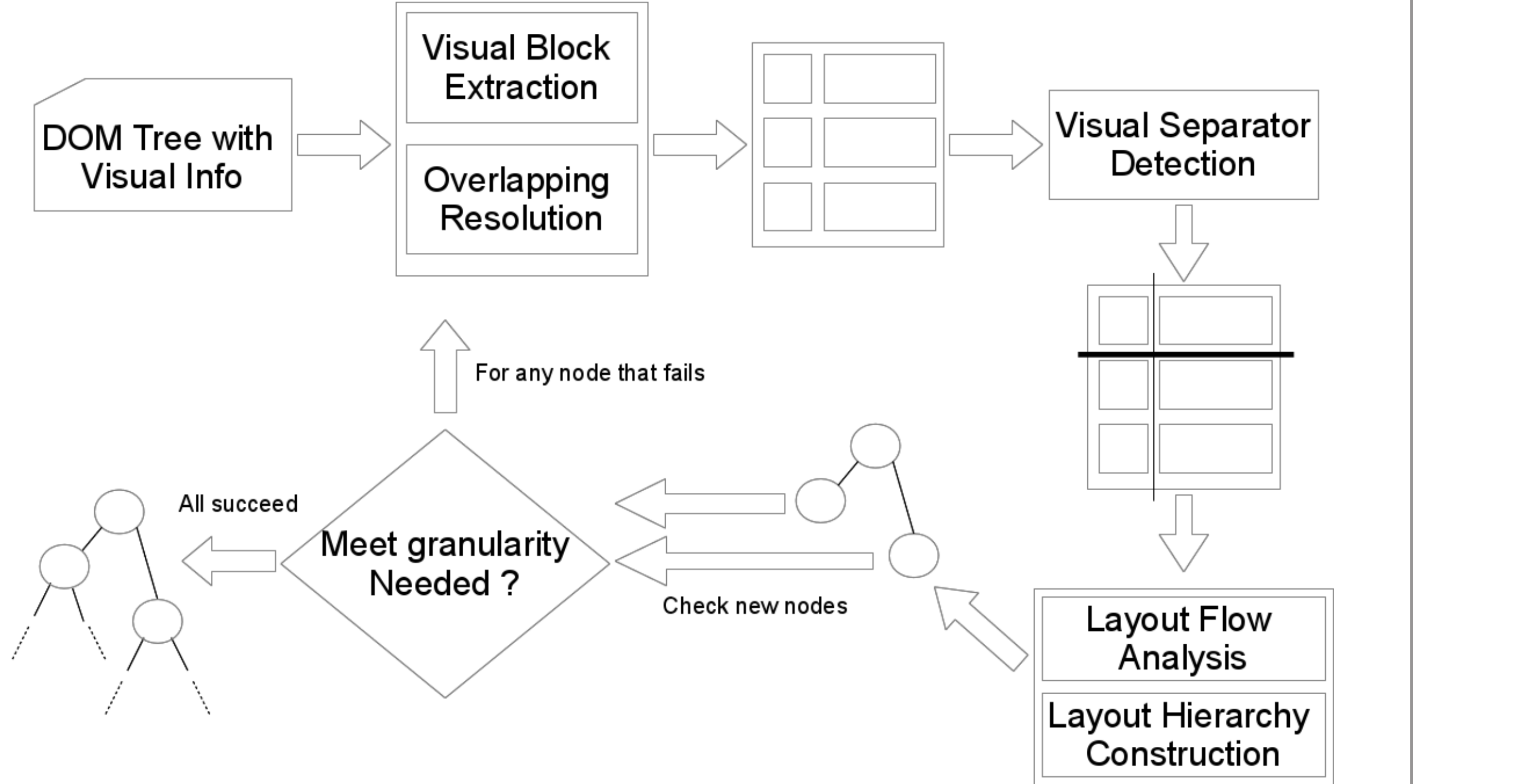
HTML Tag with visual cues added

```
<div id="item"
background-color="rgba(255,255,255,0)"
font-style="Arial"
font-size="12px"
left="64"
top="0"
width="200"
height="80"
>
...
</div>
```

**Output:**

- Visual Blocks Hierarchy  
A XML file denoting the hierarchy of Visual Blocks
- Feedback information  
User can adjust block geometry, add or delete blocks. This information is stored in a database for later use

## Segmentation Process



**DOM Tree with Visual Info:** Is the DOM obtained from a rendering engine, after finish processing the web page

**Visual Block Extraction:** Is the implementation of the segmentation algorithm. It follows the general insight described by the author

**Overlapping Resolution:** It's objective is to resolve any overlapping between blocks an convert them to polygons

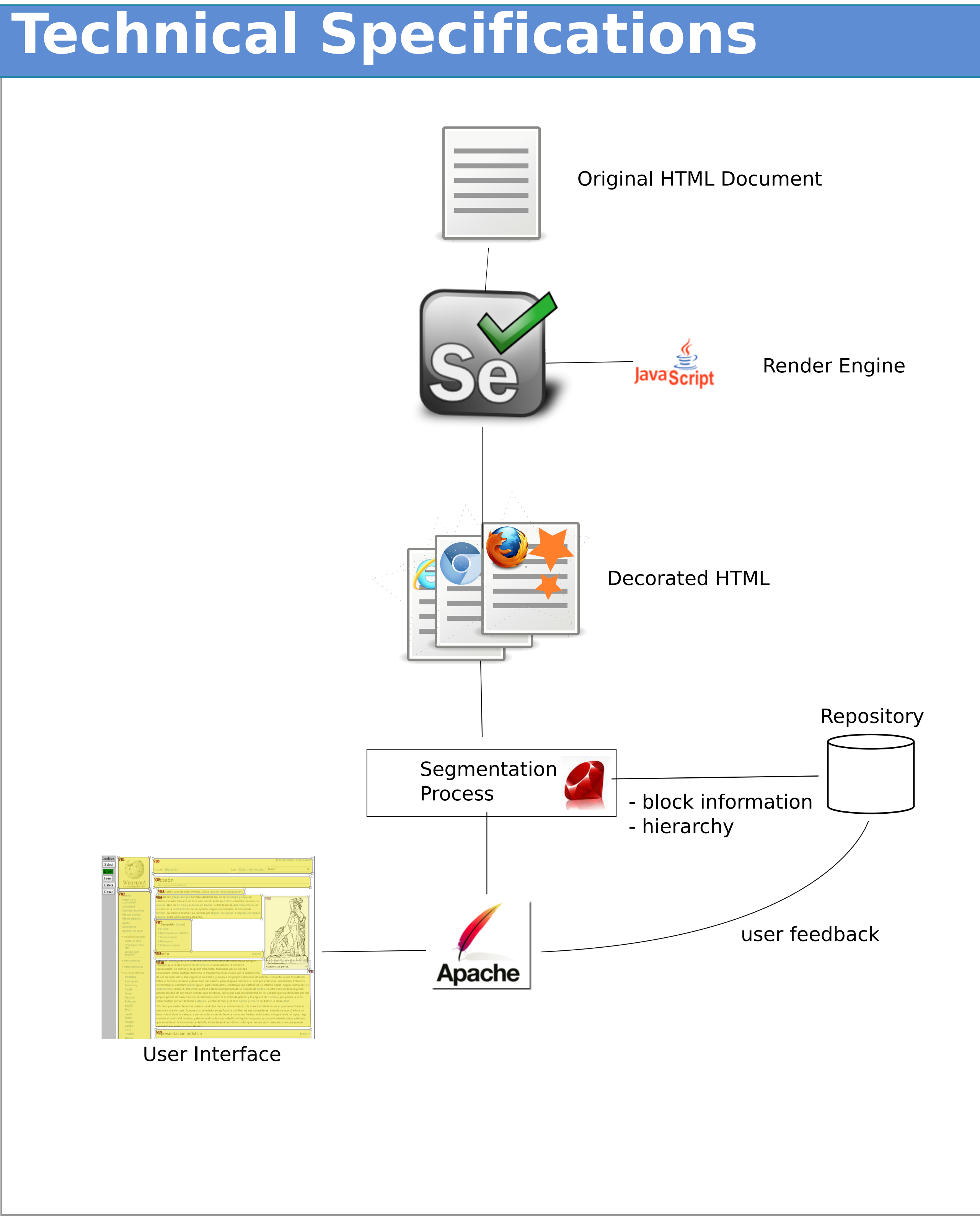
**Visual Separators Detection:** It is the process of determine the spaces or visual divisions between blocks. Here we applied computation gometry to convex polygons

**Layout Flow Analysis:** Given a hierarchy and visual separators, the latter are weighed. A higher weight higher coherence. Those with lower weights are merged to produce a more coherent block

**Layout Hierarchy Construction:** Based on predefined layout templates the hierarchy pass through a process of comparison, placing each block into a taxonomy


The segmentation algorithm is applied recursively to obtain the visual blocks and the hierarchy

It follows the insights presented by (Cai, 2003), but with the Overlapping Resolution and Layout Flow Analysis added



## User Interface

Screenshot of the tool with a Wikipedia page



Data visualization and user interaction are done though a web application. Users give the URL of the pages they want to process. The can use the basic web resource operations (add, modify, delete and list). A page for visualizing the segmentation results in a HTML5 compliant browser is available. The users can interact with the visual blocks detected: move them, tune them, add and delete new ones. This feedback information is stored for later use.

## Future Directions

This tool aims to be part of the tools developed in the framework of the SCAPE project. Our paper points out the issues of efficiently segmenting web pages and improving the quality and the relevance of the segmentation.

To address this issue, we propose an hybrid approach that use both the DOM analysis and image segmentation. We look forward to improve more significantly the precision of the segmentation and include those insights in all aspects of an archive, indexes for example.

- We look forward to improve more significantly the precision of the segmentation and include those insights in all aspects of an archive, indexes for example.
- Another on-going work is to experiment our algorithm with a bigger set of web pages, not only test cases. We intend to use the feedback from the users as a input to a Knowledge Base which would help to improve the Visual Block Extraction Phase, therefore the quality of the archive itself.
- Beside this, we are would like to test the hybrid approach to include information that usually with DOM approaches only are not reach, such as flash movies.

### References

E. Bruno, N. Faessel, H. Glotin, J. Le Maitre, and M. Scholl. Indexing by permeability in block structured web pages. In Proc. of the 9th ACM symposium on Document engineering, DocEng '09, pages 70-73, New York, NY, USA, 2009. ACM.

D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. Extracting content structure for web pages based on visual representation. In Fifth Asia Pacific Web Conference (APWeb'03), 2003.

B. Kruatrachue, N. Moongfangklang, and K. Siriboon. Fast document segmentation using contour and x-y cut technique. In C. Ardil, editor, WEC (5), pages 27-29. Enformatika, Canakkale, Turkey, 2005.

M. B. Saad and S. Gançarski. Using visual pages c analysis for optimizing web archiving. In Proc. of the 2010 EDBT/ICDT Workshops, EDBT '10, pages 43:1-43:7, New York, NY, USA, 2010. ACM.

X. Yang and Y. Shi. Enhanced gestalt theory guided web page segmentation for mobile browsing. In Proc. of the International Joint Conference on Web Intelligence and Intelligent Agent Technology, pages 46-49, Washington, DC, USA, 2009.