

# Merge Semi Estruturado HTML

Pedro Henrique - phsm

# Foco da gramática

Os pontos de atuação do merge semi estruturado na gramática são em atributos de tags e em quaisquer tipos de elementos, como tags, comentários, scriptlets, CDATA, referências de DTDs, XML Schemas e texto.

# Principais repositórios minerados

<https://github.com/lestencrypt/website>

<https://github.com/PaystackHQ/website>

<https://github.com/pattern-lab/website>

<https://github.com/msgpack/website>

<https://github.com/lerna/website>

<https://github.com/pallets/website>

<https://github.com/syncthing/website>

<https://github.com/mathletes/website>

<https://github.com/freenet/website>

<https://github.com/churchio/website>

<https://github.com/h5bp/html5-boilerplate>

# Arquivos HTML

Quantidade de arquivos  
com conflitos

NE -> Não estruturado

SE -> Semi estruturado

EQ -> Não e semi  
estruturado

Repositório	Merges	Conflitos NE	Conflitos SE	Conflitos EQ
Letsencrypt	549	0	0	0
PaystackHQ	52	2	2	2
Pattern-lab	33	0	0	0
Msgpack	211	1	1	1
Lerna	6	0	0	0
Pallets	100	0	0	0
Syncthing	63	0	0	0
Mathletes	361	1	4	1
Freenet	187	0	0	0
Churchio	23	1	1	1
Html5-bp	236	4	4	4

# Conflitos

Quantidade de conflitos

NE -> Não estruturado

SE -> Semi estruturado

EQ -> Não e semi estruturado

Repositório	Conflitos NE	Conflitos SE	Conflitos EQ
Letsencrypt	0	0	0
PaystackHQ	2	2	2
Pattern-lab	0	0	0
Msgpack	1	1	1
Lerna	0	0	0
Pallets	0	0	0
Syncthing	0	0	0
Mathletes	1	104	0
Freenet	0	0	0
Churchio	1	1	1
Html5-bp	4	4	4

# PaystackHQ - Conflitos em ambos

<<<<<< MINE

```
<li class="global-nav__item global-nav__item--has-dropdown">
  <a href="#">
    Product <i class="icon icon--navigation-chevron"></i>
  </a>
  <ul class="global-dropdown-menu">
    ... ..
  </ul>
</li>
```

=====

```
<li class="global-nav__item"><a href="product.html">Product</a></li>
```

>>>>>> YOURS

LEFT

```
<div class="global-nav__right">
  <ul class="global-nav__menu">
    <li class="global-nav__item global-nav__item--has-dropdown">
      <a href="#">
        Product <i class="icon icon--navigation-chevron"></i>
      </a>
      <ul class="global-dropdown-menu">
        <li class="global-dropdown-menu__item">
          <a href="#">Payments</a>
        </li>
        <li class="global-dropdown-menu__divider"></li>
        <li class="global-dropdown-menu__item">
          <a href="#">Business Tools</a>
        </li>
        <li class="global-dropdown-menu__divider"></li>
        <li class="global-dropdown-menu__item">
          <a href="#">Developer Tools</a>
        </li>
      </ul>
    </li>
    <li class="global-nav__item"><a href="#">Pricing</a></li>
```

BASE

```
<div class="global-nav__right">
  <ul class="global-nav__menu">
    <li class="global-nav__item">
      <a href="#">Product</a>
      <ul class="global-dropdown-menu">
        <li class="global-dropdown-menu__item">
          <a href="#">Payments</a>
        </li>
        <li class="global-dropdown-menu__divider"></li>
        <li class="global-dropdown-menu__item">
          <a href="#">Business Tools</a>
        </li>
      </ul>
    </li>
    <li class="global-nav__item"><a href="#">Pricing</a></li>
```

TP

RIGHT

```
<div class="global-nav__right">
  <ul class="global-nav__menu">
    <li class="global-nav__item"><a href="product.html">Product</a></li>
    <li class="global-nav__item"><a href="#">Pricing</a></li>
```

Motivo do conflito em ambos é claro, pois o conteúdo do LEFT e RIGHT diferem da BASE para o não estruturado, e a tag `<li>` que contém as diferenças do LEFT e RIGHT possui o mesmo tipo para o merge semi estruturado. O merge semi estruturado pode ser modificado para levar em conta o tipo de acordo com os atributos e permitir merge nesses casos, mas deve haver cautela, pois nesse caso a informação seria duplicada.

# PaystackHQ - EQ

<<<<<< MINE



=====



>>>>>> YOURS



## LEFT

```
<div class="grid__column grid__column--6">
  <div class="features-illustration">
    
  </div>
</div>
```

## BASE

```
<div class="grid__column grid__column--6">
  <div class="features-illustration">
    
  </div>
</div>
```

## RIGHT

```
<div class="grid__column grid__column--6">
  <div class="features-illustration">
    
  </div>
</div>
```

O merge semi estruturado não é capaz de resolver o valor de um atributo, pois isso é terminal, e como as linhas em do LEFT, RIGHT e BASE são todas diferentes, o não estruturado também não consegue resolver e reporta conflito.

# Msgpack - EQ

<<<<<< MINE

<p class="getstart\_links">

<!--<a href="http://redmine.msgpack.org/projects/msgpack/wiki/QuickStartPerl">Quick Start</a-->

<span style="margin-left:0;">Reference:</span>

<a href="http://search.cpan.org/~tokuhirom/AnyEvent-MPRPC/">RPC</a>

<a href="http://search.cpan.org/~tokuhirom/Data-MessagePack/">Serialization</a>

=====

<pre>cpan&gt; install Data::MessagePack

cpan&gt; install AnyEvent::MPRPC</pre>

<p>

See <a href="http://search.cpan.org/~tokuhirom/Data-MessagePack/">Data::MessagePack CPAN page</a> for detail.

<a href="http://search.cpan.org/~tokuhirom/AnyEvent-MPRPC/">AnyEvent::MPRPC</a> is a RPC implementation on

top of AnyEvent.

>>>>>> YOURS

LEFT

```
<div class="box3">
  <h3>Perl</h3>
  <p class="getstart_links">
    <!--<a href="http://redmine.msgpack.org/projects/msgpack/wiki/QuickStartPerl">Quick Start</a>-->
    <span style="margin-left:0;">Reference:</span>
    <a href="http://search.cpan.org/~tokuhirom/AnyEvent-MPRPC/">RPC</a>
    <a href="http://search.cpan.org/~tokuhirom/Data-MessagePack/">Serialization</a>
  </p>
  <pre>&gt; cpan Data::MessagePack
&gt; cpan AnyEvent::MPRPC</pre>
</div>
```

BASE

ESSA DIV NÃO EXISTE NA BASE (ELAS APARECEM DENTRO DA MESMA DIV PAI)

RIGHT

```
<div class="box3">
  <h3>Perl</h3>
  <pre>cpan&gt; install Data::MessagePack
cpan&gt; install AnyEvent::MPRPC</pre>
  <p>
    See <a href="http://search.cpan.org/~tokuhirom/Data-MessagePack/">Data::MessagePack CPAN page</a> for detail.
    <a href="http://search.cpan.org/~tokuhirom/AnyEvent-MPRPC/">AnyEvent::MPRPC</a> is a RPC implementation on top of AnyEvent.
  </p>
</div>
```

# Mathletes - NE

<<<<<< MINE

```
<meta name="viewport" content="width=device-width, initial-scale=1.0, maximum-scale=1.0, user-scalable=no">
```

```
<link href="static/layout/styles/layout.css" rel="stylesheet" type="text/css" media="all">
```

```
<link rel="icon" type="image/x-icon" href="static/images/demo/favicon.ico">
```

=====

```
<meta name="viewport" content="width=device-width, initial-scale=1.0, maximum-scale=1.0, user-scalable=no" />
```

```
<link href="static/layout/styles/layout.css" rel="stylesheet" type="text/css" media="all" />
```

```
<link href="static/layout/styles/searchlayout.css" rel="stylesheet" type="text/css" media="all" />
```

>>>>>> YOURS

## LEFT

```

<head>
<title>Mathletes</title>
<meta charset="utf-8">
<meta name="viewport" content="width=device-width, initial-scale=1.0, maximum-scale=1.0, user-scalable=no">
<link href="static/layout/styles/layout.css" rel="stylesheet" type="text/css" media="all">
<link rel="icon" type="image/x-icon" href="static/images/demo/favicon.ico">
</head>

```

Esse conflito é causado por uma falha na gramática que identifica tags vazias `<tag attrs />` (prefixo `ettag->`) e `<tag attrs >` (prefixo `ettag->`) com prefixos diferentes, a correção desse problema permitirá que o merge semi estruturado ocorra sem erros.

## BASE

```

<head>
<title>Mathletes</title>
<meta charset="utf-8">
<meta name="viewport" content="width=device-width, initial-scale=1.0, maximum-scale=1.0, user-scalable=no">
<link href="static/layout/styles/layout.css" rel="stylesheet" type="text/css" media="all">
</head>

```

## RIGHT

```

<head>
<title>Mathletes</title>
<meta charset="utf-8">
<meta name="viewport" content="width=device-width, initial-scale=1.0, maximum-scale=1.0, user-scalable=no" />
<link href="static/layout/styles/layout.css" rel="stylesheet" type="text/css" media="all" />
<link href="static/layout/styles/searchlayout.css" rel="stylesheet" type="text/css" media="all" />
</head>

```

# Mathletes - SE

- 104 Conflitos do tipos:

```
<<<<<< MINE  
"fl_left"  
=====  
>>>>>> YOURS
```

LEFT

```
<div class="wrapper row0">  
  <div id="topbar" class="hoc clear">  
    <div class="fl_left">  
      <ul>
```

BASE

```
<div class="wrapper row0">  
  <div id="topbar" class="hoc clear">  
    <div class="fl_left">  
      <ul>
```

RIGHT

```
<div class="wrapper row0">  
  <div id="topbar" class="hoc clear">  
    <div class="fl_left">  
      <ul>
```

Caso com bug, o merge semi estruturado não deveria reportar um conflito, pois a estrutura dos htmls não muda na região do conflito, que provavelmente foi causado por uma falha na gramática.

Esse problema aconteceu apenas neste projeto, mas para vários valores de atributo e em quatro revisões do mesmo arquivo.

# Churchio - EQ

<<<<<< MINE

Copyright &copy; 2014 <a href="http://timmorgan.org">Tim Morgan</a> |  
<a href="/open-source-church-software">List of Open Source Church Software</a>

=====

Copyright &copy; 2014 <a href="http://timmorgan.org/" onclick="ga('send', 'event' , 'Outbound Click' ,  
this.href, " ); setTimeout(function () { document.location.href=link.href; }, 150); return false;">Tim  
Morgan</a>

>>>>>> YOURS



## LEFT

```
<div class="row">
  <div class="col-lg-12 text-center">
    Copyright &copy; 2014 <a href="http://timmorgan.org">Tim Morgan</a> |
    <a href="/open-source-church-software">List of Open Source Church Software</a>
  </div>
</div>
```

Link em href é terminal, provocando o conflito pois o LEFT e RIGHT possuem valores diferentes.

## BASE

```
<div class="row">
  <div class="col-lg-12 text-center">
    Copyright &copy; 2014 <a href="timmorgan.org">Tim Morgan</a>
  </div>
</div>
```

## RIGHT

```
<div class="row">
  <div class="col-lg-12 text-center">
    Copyright &copy; 2014 <a href="http://timmorgan.org/" onclick="ga('send', 'event', 'Outbound Click', this.href, " "); setTimeout(function
() { document.location.href=link.href; }, 150); return false;">Tim Morgan</a>
  </div>
</div>
```

# Html5-boilerplate - EQ

```
<<<<<< MINE
```

```
<!-- All JavaScript at the bottom, except for Modernizr. -->
```

```
<script src="js/modernizr-1.5.min.js"></script>
```

```
=====
```

```
<!-- All JavaScript at the bottom, except for.. -->
```

```
<script>
```

```
// http://paulirish.com/2009/avoiding-the-fouc-v3/ &
```

```
http://remysharp.com/2009/01/07/html5-enabling-script
```

```
// alternatively using Modernizr does both of these.
```

```
(function(B,C){B[C]=B[C].replace(/\bno-js\b/,'js');if(!/*@cc_on!@*/0) ... .. ment,'className');
```

```
</script>
```

```
>>>>>> YOURS
```

LEFT

```
<link rel="stylesheet" media="handheld" href="css/mobile.css">
```

```
<!-- All JavaScript at the bottom, except for Modernizr. -->
```

```
<script src="js/modernizr-1.5.min.js"></script>
```

```
</head>
```

BASE

```
<link rel="stylesheet" media="handheld" href="css/mobile.css">
```

```
<!-- All JavaScript at the bottom, except for.. -->
```

```
<script>
```

```
// http://paulirish.com/2009/avoiding-the-fouc-v3/ & http://remysharp.com/2009/01/07/html5-enabling-script
```

```
// alternatively using Modernizr does both of these.
```

```
(function(B,C){B[C]=B[C].replace(/\bno-js\b/,'js');if(!/*@cc_on!@*/0) ... .. mentElement,'className');
```

```
</script>
```

```
</head>
```

RIGHT

```
<link rel="stylesheet" media="handheld" href="css/mobile.css">
```

```
<!-- All JavaScript at the bottom, except for.. -->
```

```
<script>
```

```
// http://paulirish.com/2009/avoiding-the-fouc-v3/ & http://remysharp.com/2009/01/07/html5-enabling-script
```

```
// alternatively using Modernizr does both of these.
```

```
(function(B,C){B[C]=B[C].replace(/\bno-js\b/,'js');if(!/*@cc_on!@*/0) ... .. mentElement,'className');
```

```
</script>
```

```
</head>
```

Caso estranho, houve conflito no merge não estruturado e no semi estruturado, apesar do BASE e RIGHT serem idênticos.

# Html5-boilerplate - EQ

<<<<<< MINE

```
<script src="//ajax.googleapis.com/ajax/libs/jquery/1.4.2/jquery.min.js"></script>
```

```
<script>!window.jQuery && document.write(unescape('%3Cscript src="js/jquery-1.4.2.min.js"%3E%3C/script%3E'))</script>
```

```
<script src="js/plugins.js?v=1"></script>
```

```
<script src="js/script.js?v=1"></script>
```

=====

```
<script src="http://ajax.googleapis.com/ajax/libs/jquery/1.4.2/jquery.js"></script>
```

```
<script>!window.jQuery && document.write('<script src="js/jquery-1.4.2.js"><\script>')</script>
```

```
<!-- scripts concatenated and minified via ant
```

```
    add new scripts to target#scripts > concat > fileset
```

```
don't delete these comments or you'll bust the build script :)
```

```
-->
```

```
<script src="js/plugins.js"></script>
```

```
<script src="js/script.js"></script>
```

```
<!-- end concatenated and minified scripts-->
```

>>>>>> YOURS

LEFT

```
<script src="//ajax.googleapis.com/ajax/libs/jquery/1.4.2/jquery.min.js"></script>
```

```
<script>!window.jQuery && document.write(unescape('%3Cscript src="js/jquery-1.4.2.min.js"%3E%3C/script%3E'))</script>
```

```
<script src="js/plugins.js?v=1"></script>
```

```
<script src="js/script.js?v=1"></script>
```

BASE

```
<script src="http://ajax.googleapis.com/ajax/libs/jquery/1.4.2/jquery.min.js"></script>
```

```
<script>!window.jQuery && document.write('<script src="js/jquery-1.4.2.min.js"></script>')</script>
```

```
<script src="js/plugins.js?v=1"></script>
```

```
<script src="js/script.js?v=1"></script>
```

RIGHT

```
<script src="http://ajax.googleapis.com/ajax/libs/jquery/1.4.2/jquery.js"></script>
```

```
<script>!window.jQuery && document.write('<script src="js/jquery-1.4.2.js"></script>')</script>
```

```
<!-- scripts concatenated and minified via ant
```

```
  add new scripts to target#scripts > concat > fileset
```

```
don't delete these comments or you'll bust the build script :)
```

```
-->
```

```
<script src="js/plugins.js"></script>
```

```
<script src="js/script.js"></script>
```

```
<!-- end concatenated and minified scripts-->
```

TP

O conteúdo de script é terminal portanto  
houve conflito nos dois merges.

# Html5-boilerplate - EQ

<<<<<< MINE

<link rel="stylesheet" media="handheld" href="css/handheld.css?v=2"> -->

=====

<link rel="stylesheet" media="handheld" href="css/handheld.css"> -->

>>>>>> YOURS

## LEFT

```
<link rel="stylesheet" href="css/style.css?v=2">
```

```
<!-- Uncomment if you are specifically targeting less enabled mobile browsers
```

```
<link rel="stylesheet" media="handheld" href="css/handheld.css?v=2"> -->
```

```
<!-- All JavaScript at the bottom, except for Modernizr which enables HTML5 elements & feature detects -->
```

```
<script src="js/helpers/modernizr-1.5.min.js"></script>
```

## BASE

```
<link rel="stylesheet" href="css/style.css?v=1">
```

```
<!-- Uncomment if you are specifically targeting less enabled mobile browsers
```

```
<link rel="stylesheet" media="handheld" href="css/handheld.css?v=1"> -->
```

```
<!-- All JavaScript at the bottom, except for Modernizr which enables HTML5 elements & feature detects -->
```

```
<script src="js/modernizr-1.5.min.js"></script>
```

## RIGHT

```
<link rel="stylesheet" href="css/style.css?v=1">
```

```
<!-- Uncomment if you are specifically targeting less enabled mobile browsers
```

```
<link rel="stylesheet" media="handheld" href="css/handheld.css"> -->
```

```
<!-- All JavaScript at the bottom, except for Modernizr which enables HTML5 elements & feature detects -->
```

```
<script src="js/modernizr-1.5.min.js"></script>
```

# TP

Comentários são não terminais, mas seu conteúdo não, por isso houve conflito nos dois merges.

# Html5-boilerplate - EQ

```
<<<<<<< MINE
<!doctype html>
<!-- saved from url=(0014)about:internet -->
<html lang="en" class="no-js">
=====
<!doctype html>

<!-- paulirish.com/2008/conditional-stylesheets-vs-css-hacks-answer-neither/ -->
<!--[if lt IE 7 ]> <html lang="en" class="no-js ie6"> <![endif]-->
<!--[if IE 7 ]> <html lang="en" class="no-js ie7"> <![endif]-->
<!--[if IE 8 ]> <html lang="en" class="no-js ie8"> <![endif]-->
<!--[if IE 9 ]> <html lang="en" class="no-js ie9"> <![endif]-->
<!--[if (gt IE 9)|!(IE)]><!--> <html lang="en" class="no-js"> <!--<![endif]-->
>>>>>>> YOURS
```



## LEFT

```
<!doctype html>
<!-- saved from url=(0014)about:internet -->
<html lang="en" class="no-js">
<head>
  <meta charset="utf-8">
```

## BASE

```
<!doctype html>
<html lang="en" class="no-js">
<head>
  <meta charset="utf-8">
```

## RIGHT

```
<!doctype html>

<!-- paulirish.com/2008/conditional-stylesheets-vs-css-hacks-answer-neither/ -->
<!--[if lt IE 7 ]> <html lang="en" class="no-js ie6"> <![endif]-->
<!--[if IE 7 ]>  <html lang="en" class="no-js ie7"> <![endif]-->
<!--[if IE 8 ]>  <html lang="en" class="no-js ie8"> <![endif]-->
<!--[if IE 9 ]>  <html lang="en" class="no-js ie9"> <![endif]-->
<!--[if (gt IE 9) | !(IE)]><!--> <html lang="en" class="no-js"> <!--<![endif]-->
<head>
  <meta charset="utf-8">
```

O conflito no merge semiestruturado acontece pois o conteúdo do comentário que foi criado no LEFT e no RIGHT são diferentes, o conteúdo é terminal portanto o conflito é reportado.

# Contagem

De 112 casos de conflito foram encontrados:

- 6 True positives
- 0 False negatives
- 106 False positives, sendo:
  - 104 Causados por bug da gramática em caso específico
  - 1 Causado por erro na definição de tipo na gramática
  - 1 Causado por erro desconhecido

Assumindo a correção dos problemas da gramática apenas 6 casos True positive restariam.

# Conclusão

A linguagem HTML possui estrutura muito flexível e vários detalhes que atrapalham a resolução de conflitos com o merge semi estruturado, suas tags podem não conter atributos ou elementos que as identifiquem, a ordem das mesmas é muito importante para a estrutura das páginas, o código html pode conter códigos de outras linguagens. Nos testes realizados podemos ver que mesmo assumindo a solução dos problemas da gramática ainda existem vários casos que o merge semi estruturado não é capaz de resolver, sendo causados principalmente por elementos terminais, ou seja, é necessária a correção manual. Isso torna o uso do merge semi estruturado para HTML não tão interessante quanto para outras linguagens como java, swift ou python.

# Dados para replicação

O programa que realiza o merge de arquivos HTML pode ser encontrado em:

<https://github.com/pedro00dk/jfstmerge>

A mineração dois projetos listados no início foi feita usando o projeto:

<https://github.com/guilhermejccavalcanti/GitMergesMiner>

O jFSTMerge deve ser adicionado como uma biblioteca ao GitMergeMiner para que funcione corretamente, após isso basta configurar os arquivos projects.csv e config.properties com o endereço do projeto e local de processamento de arquivos.