Frédéric Vergnaud

Centre de Sociologie de l'Innovation – Mines-ParisTech – PSL –
Institut Interdisciplinaire de l'Innovation (CNRS UMR 9217)

Le scraping de données conversationnelles avec Extractify

Les rencontres de statistique appliquée - INED
Vendredi 12 Février 2021

Qu'est-ce que le web ?

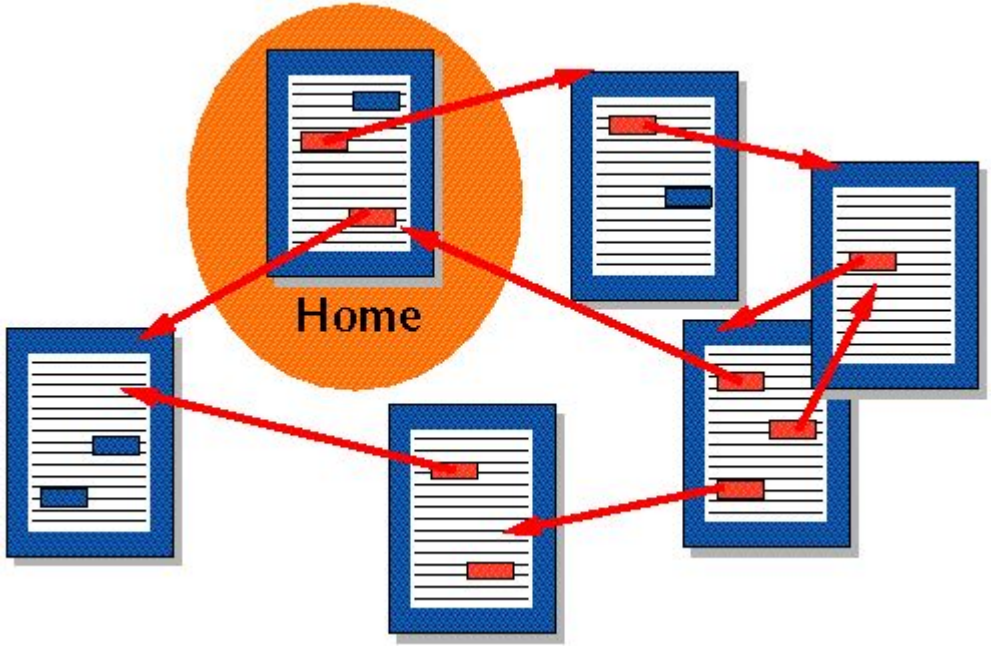
C'est un « Système hypermédia permettant d'accéder aux ressources du réseau Internet »

Qu'est-ce qu'Internet ?

C'est un « Réseau télématique international [...] résultant de l'interconnexion d'ordinateurs du monde entier utilisant un protocole commun d'échanges de données »

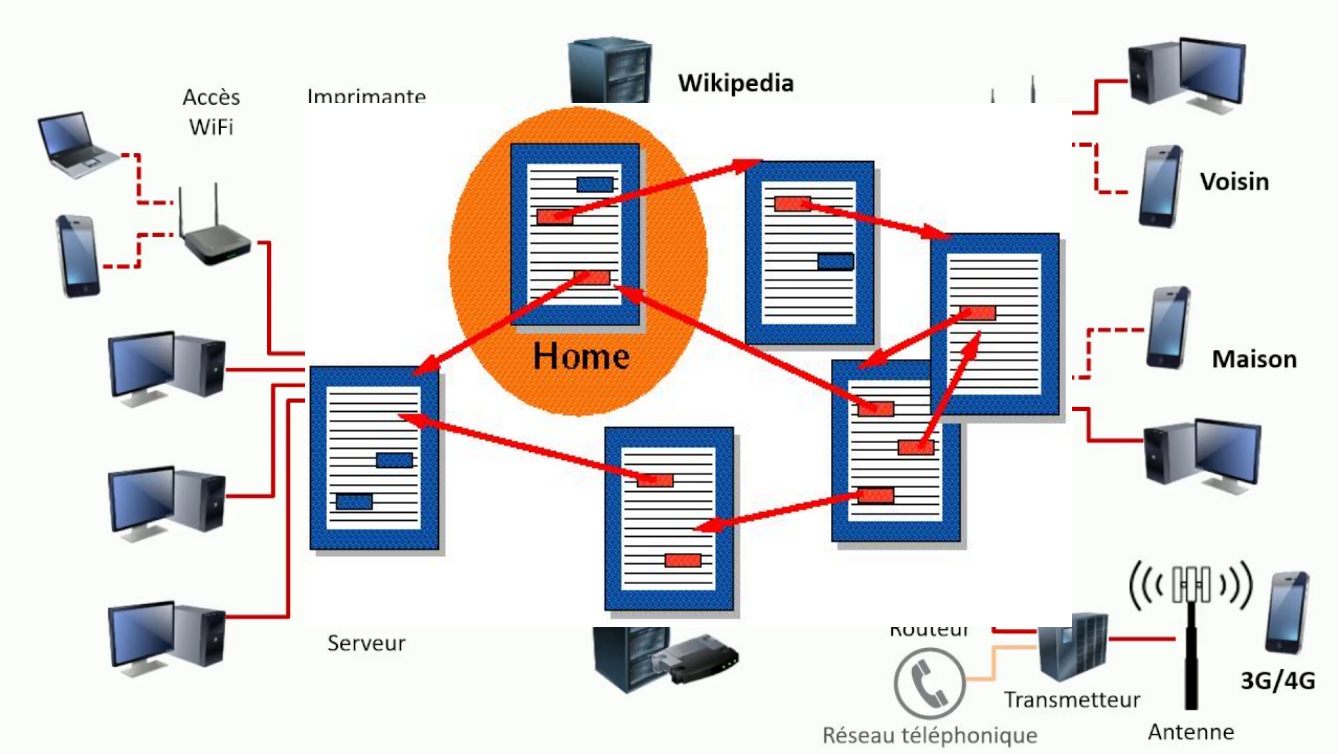
Qu'est-ce qu'un hypermédia ?

C'est un « mode d'organisation des documents informatisés caractérisé par l'existence de liens dynamiques entre ses différentes sections »



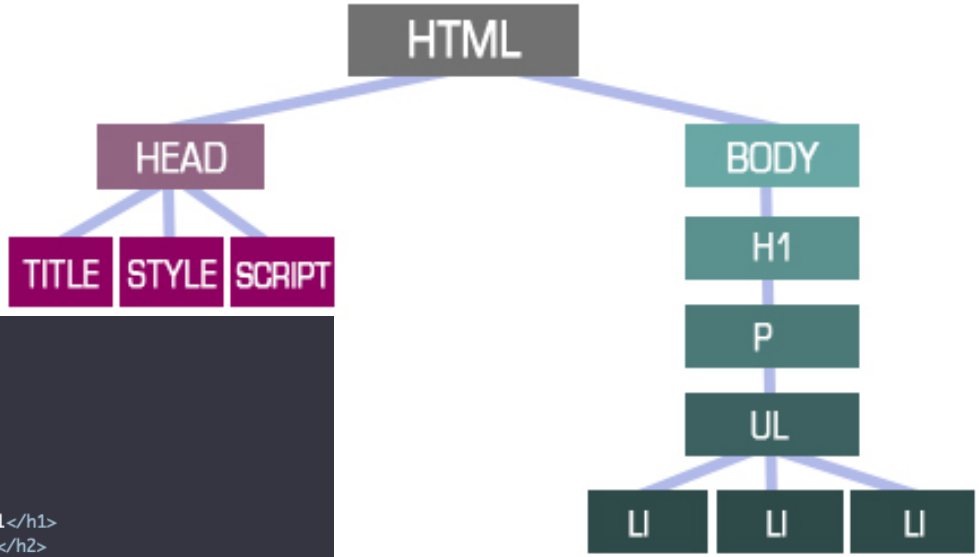
Qu'est-ce que le web ?

C'est un « Système hypermédia permettant d'accéder aux ressources du réseau Internet »



Qu'est-ce que le HTML ?

C'est un « langage de description de format de document qui se présente sous la forme d'un langage de balisage »




```
<!DOCTYPE html>
<html>
    <head>
        <title>Ma première page HTML</title>
        <meta charset="utf-8">
    </head>
    <body>
        <!--Les espaces ne seront pas affichés-->
        <h1>Mon titre                    principal</h1>
        <h2>Je suis       un       titre important</h2>

        <p>Voici
            mon
            premier
            paragraphe.</p>
        <p>Et en voilà un second !</p>
    </body>
</html>
```

Navigateur web

« Cher serveur de l'Ined, peux-tu m'afficher la page web dont l'adresse est https://www.ined.fr/ ? »

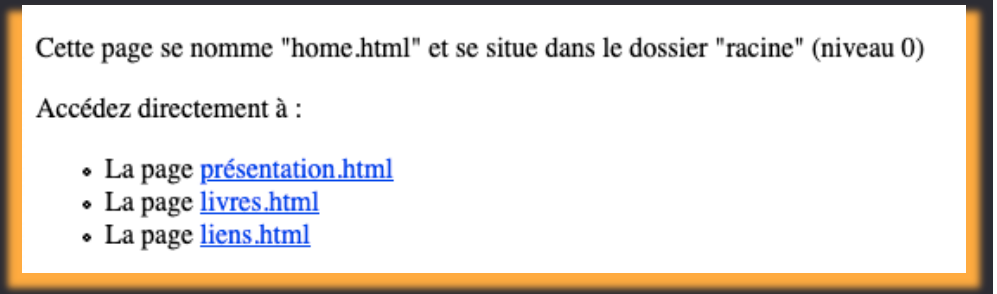
« Oh oui bien sûr cher client, la voici la voilà ! »

Serveur

Le HTML est un langage à balises

```html
<!DOCTYPE html>
<html>
    <head>
        <title>Home</title>
        <meta charset='utf-8'>
    </head>
    <body>
        <p>Cette page se nomme "home.html" et se situe
        dans le dossier "racine" (niveau 0)</p>

        <p>Accédez directement à :</p>
        <ul>
            <li>La page <a href="cours/presentation.html">presentation.html</a></li>
            <li>La page <a href="cours/livres.html">livres.html</a></li>
            <li>La page <a href="cours/code/liens.html">liens.html</a></li>
        </ul>
    </body>
</html>
```



Une balise HTML définit un élément

```
<!DOCTYPE html>
<html>
    <head>
        <title>Home</title>
        <meta charset='utf-8'>
    </head>
    <body>
        <p>Cette page se nomme "home.html" et se situe
        dans le dossier "racine" (niveau 0)</p>

        <p>Accédez directement à :</p>
        <ul>
            <li>La page <a href="cours/presentation.html">presentation.html</a></li>
            <li>La page <a href="cours/livres.html">livres.html</a></li>
            <li>La page <a href="cours/code/liens.html">liens.html</a></li>
        </ul>
    </body>
</html>
```

Cette page se nomme "home.html" et se situe dans le dossier "racine" (niveau 0)

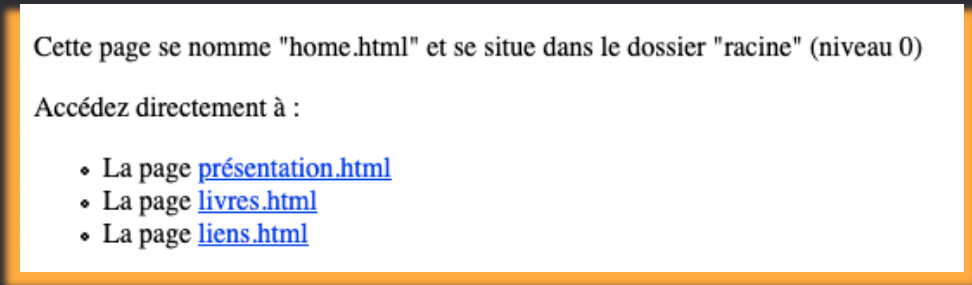
Accédez directement à :

- La page présentation.html
- La page livres.html
- La page liens.html

Un élément peut avoir des attributs. Ex: href

```
<!DOCTYPE html>
<html>
    <head>
        <title>Home</title>
        <meta charset='utf-8'>
    </head>
    <body>
        <p>Cette page se nomme "home.html" et se situe
        dans le dossier "racine" (niveau 0)</p>

        <p>Accédez directement à :</p>
        <ul>
            <li>La page <a href="cours/presentation.html">presentation.html</a></li>
            <li>La page <a href="cours/livres.html">livres.html</a></li>
            <li>La page <a href="cours/code/liens.html">liens.html</a></li>
        </ul>
    </body>
</html>
```



Un élément peut avoir des attributs. Ex: src

```
<!DOCTYPE html>
<html>
    <head>
        <title>Cours HTML et CSS</title>
        <meta charset= "utf-8">
        <link rel="stylesheet" href="cours.css">
    </head>

    <body>
        <div class="conteneur">
            <img class="sun" src="sunset.png" alt="Coucher de soleil">
        </div>
        <div class="conteneur">
            <img class="sun" src="sunset.png" alt="Coucher de soleil">
        </div>
        <img class="sun bloc" src="sunset.png" alt="Coucher de soleil">
        <img class="sun bloc" src="sunset.png" alt="Coucher de soleil">
    </body>
</html>
```

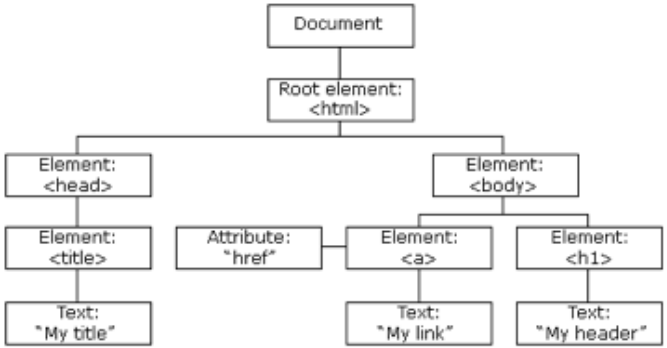
Le DOM (Document Object Model) : une arborescence d'éléments HTML

```html
<ul>
  <li>La page
    <a href="cours/présentation.html">présentation.html</a>
  </li>
  <li>La page
    <a href="cours/livres.html">livres.html</a>
  </li>
  <li>La page
    <a href="cours/liens.html">liens.html</a>
  </li>
</ul>
```

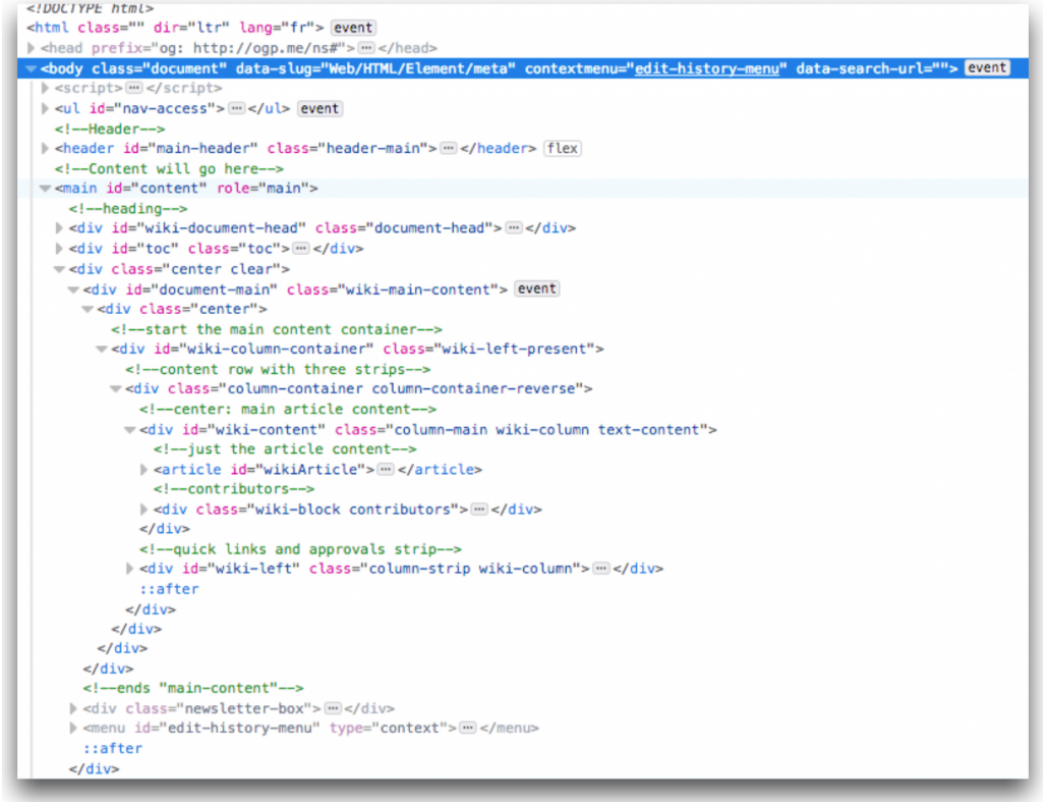
Le document HTML

```html
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="utf-8">
    <title>My Title</title>
</head>
<body>
    <a href="http://www.monsite.fr">My link</a>
    <h1>My header</h1>
</body>
</html>
```

Le DOM

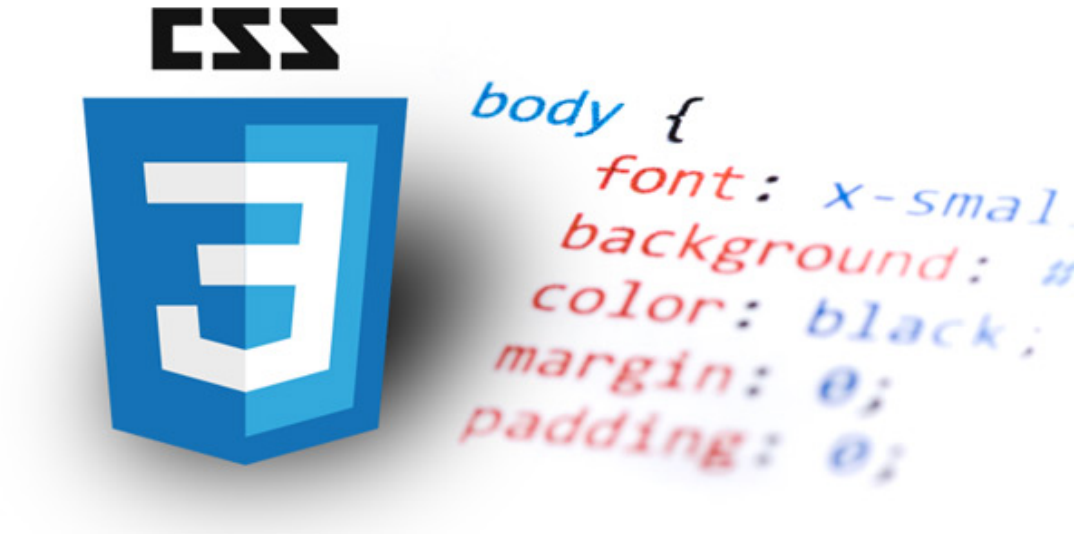


Une page HTML « normale »



Qu'est-ce que le CSS (Cascading Style Sheets – Feuilles de style en cascade) ?

C'est un « langage qui décrit la présentation des pages HTML »



2 sélecteurs CSS simples : p et body

HTML

```
<!DOCTYPE html>
<html>
  <head>
    <title>Home</title>
    <meta charset='utf-8'>
  </head>
  <body>
    <p>Cette page se nomme "home.html" et se situe
      dans le dossier "racine" (niveau 0)</p>
    <p>Accédez directement à :</p>
    <ul>
      <li>La page <a
href="cours/présentation.html">présentation.html</a></li>
```

CSS 14 unsaved changes

```
p {
  color: blue;
}

body {
  background-color: FloralWhite;
  text-align: center;
}

#bonjour {
  color:crimson;
  font-family: Verdana;
  background-color: Pink;
```

Cette page se nomme "home.html" et se situe dans le dossier "racine" (niveau 0)

Accédez directement à :

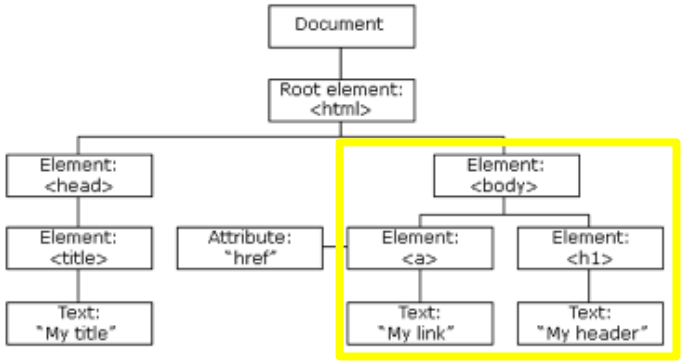
- La page présentation.html
- La page livres.html
- La page liens.html

Le DOM (Document Object Model) : une arborescence d'éléments HTML

Le document HTML

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="utf-8">
    <title>My Title</title>
</head>
<body>
    <a href="http://www.monsite.fr">My link</a>
    <h1>My header</h1>
</body>
</html>
```

Le DOM



2 sélecteurs CSS simples : p et body

HTML

```
<!DOCTYPE html>
<html>
  <head>
    <title>Home</title>
    <meta charset='utf-8'>
  </head>
  <body>
    <p>Cette page se nomme "home.html" et se situe
      dans le dossier "racine" (niveau 0)</p>
    <p>Accédez directement à :</p>
    <ul>
      <li>La page <a
href="cours/présentation.html">présentation.html</a></li>
```

CSS 14 unsaved changes

```
p {
  color: blue;
}

body {
  background-color: FloralWhite;
  text-align: center;
}

#bonjour {
  color:crimson;
  font-family: Verdana;
  background-color: Pink;
```

Cette page se nomme "home.html" et se situe dans le dossier "racine" (niveau 0)

Accédez directement à :

- La page présentation.html
- La page livres.html
- La page liens.html

2 sélecteurs CSS simples d'attributs : class (.) et id (#)

HTML

```
<html>
  <head>
    <title>Home</title>
    <meta charset='utf-8'>
  </head>
  <body>
    <p id="bonjour">Cette page se nomme "home.html" et se situe
      dans le dossier "racine" (niveau 0)</p>
    <p class="paragraphe">Accédez directement à :</p>
    <ul>
      <li>La page <a href="cours/présentation.html">présentation.html</a></li>
      <li>La page <a href="cours/livres.html">livres.html</a></li>
      <li>La page <a href="cours/liens.html">liens.html</a></li>
```

CSS

```
#bonjour {
  color:crimson;
  font-family: Verdana;
  background-color: Pink;
}

.paragraphe {
  color: green;
  font-weight: 700;
  font-family: "Arial black";
}
```

Cette page se nomme "home.html" et se situe dans le dossier "racine" (niveau 0)

Accédez directement à :

- La page présentation.html
- La page livres.html
- La page liens.html

Un grand nombre de sélecteurs CSS 3 ...

Sélecteur	Description	Version CSS
E	Sélectionne tous les élément de type E	1
E .maclasse	Sélectionne les éléments E possédant un attribut class="maclasse"	1
E #monid	Sélectionne un élément E possédant un attribut id="monid"	1
E:link	Sélectionne tout élément E représentant l'ancre d'un lien non visité jusqu'à présent	1
E:visited	Sélectionne tout élément E représentant l'ancre d'un lien déjà visité	1
E, F	Sélectionne tous les éléments de type E et de type F	1
E F	Sélectionne tous les éléments F à l'intérieur des éléments E	1
E:active	Sélectionne un élément E qui est actuellement cliqué	1
E::first-letter	Sélectionne la première lettre de tout élément E	1
E::first-line	Sélectionne la première ligne (dans le résultat final) de tout élément E	1
*	Sélectionne tous les éléments	2
E > F	Sélectionne les éléments F enfants directs des éléments E	2
E + F	Sélectionne tout élément F placé directement après un élément E	2
E[foo]	Sélectionne tout élément E possédant un attribut foo	2
E[foo="bar"]	Sélectionne tout élément E possédant un attribut foo dont la valeur est exactement « bar »	2
E[foo~="bar"]	Sélectionne tout élément E possédant un attribut foo dont la valeur contient distinctement « bar » (c'est-à-dire dont la valeur contient le mot « bar » séparé du reste par des espaces)	2
E[foo|="en"]	Sélectionne tout élément E possédant un attribut foo dont la valeur commence par « en » séparé du reste par un tiret (ou hyphen en anglais)	2
E:hover	Sélectionne un élément E qui se situe sous le curseur de la souris	2
E:focus	Sélectionne un élément E qui a le focus	2
E:first-child	Sélectionne tout élément E étant le premier enfant de son parent	2
E:lang(fr)	Sélectionne tout élément E dont l'attribut langage possède la valeur « fr »	2
E::after	Ce sélecteur permet d'intégrer du contenu après un élément E	2
E::before	Ce sélecteur permet d'intégrer du contenu avant un élément E	2
E ~ F	Sélectionne tout élément F placé après un élément E dans la page	3
E[foo^="bar"]	Sélectionne tout élément E possédant un attribut foo dont la valeur commence exactement par « bar »	3
E[foo$="bar"]	Sélectionne tout élément E possédant un attribut foo dont la valeur se termine exactement par « bar »	3
E[foo*="bar"]	Sélectionne tout élément E possédant un attribut foo dont la valeur contient la valeur « bar »	3
E:target	Sélectionne un élément E contenant une ancre qui vient d'être cliquée à partir d'un lien ancre	3
E:enabled	Sélectionne tout élément E avec lequel l'utilisateur peut interagir et qui est activé	3
E:disabled	Sélectionne tout élément E avec lequel l'utilisateur peut interagir et qui est desactivé	3
E:root	Sélectionne un élément E racine du document	3
E:empty	Sélectionne tout élément E qui ne possède pas d'enfant (ni de noeud de type texte)	3
E:nth-child(n)	Sélectionne tout élément E étant le n-ième enfant de son parent	3
E:nth-last-child(n)	Sélectionne tout élément E étant le n-ième enfant de son parent en comptant les enfants à partir du dernier	3
E:checked	Sélectionne tout élément E de type input coché au sens large (checked ou selected)	3
E:last-child	Sélectionne tout élément E étant le dernier enfant de son parent	3
E:only-child	Sélectionne tout élément E qui est le seul enfant de son parent	3
E:nth-of-type(n)	Sélectionne tout élément E étant le n-ième enfant d'un certain type par rapport à son parent	3
E:nth-last-of-type(n)	Sélectionne tout élément E étant le n-ième enfant d'un certain type par rapport à son parent en comptant à partir de la fin	3
E:first-of-type	Sélectionne tout élément E premier enfant de son type par rapport à son parent	3
E:last-of-type	Sélectionne tout élément E dernier enfant de son type par rapport à son parent	3
E:only-of-type	Sélectionne tout élément E seul enfant de son type par rapport à son parent	3

… et dans les futures versions de CSS

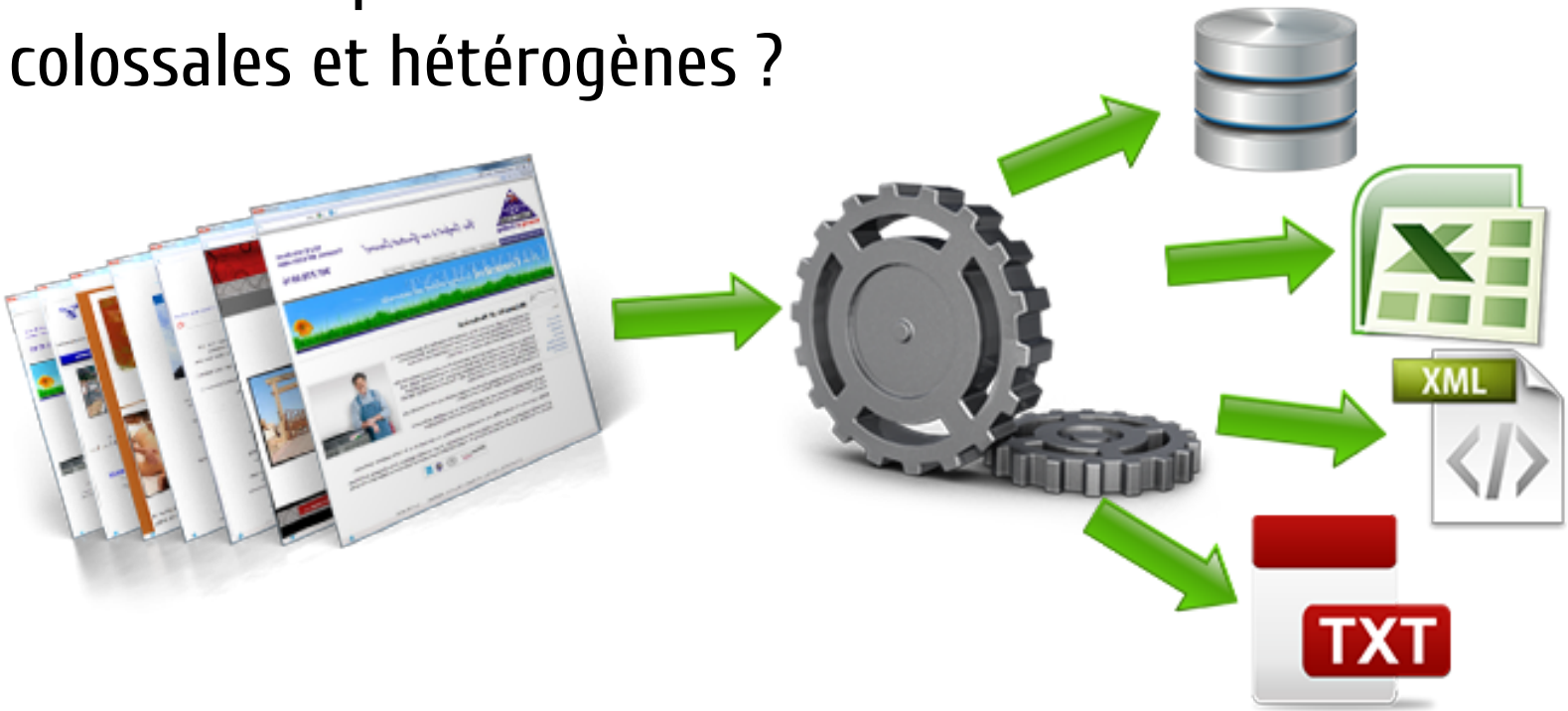
E:read-write	Sélectionne tout élément E de type input avec lequel l'utilisateur peut interagir (comme un champ dans lequel il peut écrire par exemple)	3-UI/4
E:read-only	Sélectionne tout élément E de type input avec lequel l'utilisateur ne peut pas interagir (éléments possédant un attribut disabled par exemple)	3-UI/4
E:placeholder-shown	Sélectionne tout élément E qui affiche actuellement la valeur de son attribut placeholder	3-UI/4
E:default	Sélectionne un élément E dans une liste ou un groupe qui est l'élément défini par défaut	3-UI/4
E:valid	Sélectionne tout élément E de type input dont la valeur est évaluée comme valide (dont la valeur possède une forme correspondant à ce qui est attendu)	3-UI/4
E:invalid	Sélectionne tout élément E de type input dont la valeur est évaluée comme invalide (valeur ne correspondant pas à ce qui est attendu)	3-UI/4
E:in-range	Sélectionne tout élément E de type input dont la valeur fournie se situe dans une fourchette de valeurs prédéfinies	3-UI/4
E:out-of-range	Sélectionne tout élément E de type input dont la valeur fournie se situe en dehors d'une certaine fourchette de valeurs prédéfinies	3-UI/4
E:required	Sélectionne tout élément E de type input dont la valeur doit être renseignée (élément possédant un attribut required)	3-UI/4
E:optional	Sélectionne tout élément E de type input dont la valeur ne doit pas obligatoirement être renseignée	3-UI/4
E:not(E1, .c1)	Sélectionne tout élément E qui n'est pas de type E1 et qui ne possède pas d'attribut class= »c1"	3/4
E::selection	Permet de sélectionner une partie active d'un élément E déjà sélectionnée par l'utilisateur	3/4
E >> F	Sélectionne tous les éléments F à l'intérieur des éléments E. Ce sélecteur est strictement équivalent au précédent mais représente la nouvelle syntaxe de celui-ci qui sera rendue officielle avec le CSS4.	4
F || E	Sélectionne un élément E qui représente une cellule dans un tableau ou dans une grille appartenant à une colonne représentée par un élément F	4
E[foo="bar" i]	Sélectionne tout élément E possédant un attribut foo dont la valeur est exactement « bar » sans tenir compte de la casse (l'utilisation de majuscules / minuscules). Le « i » représente ici « case insensitive » c'est-à-dire « insensible à la casse » en français	4
E:any-link	Sélectionne tout élément E étant l'ancre d'un lien, que celui-ci ait été cliqué ou non. Utiliser cette pseudo-classe est équivalent à utiliser :link et :visited	4
E:past	Sélectionne un élément E dont l'affichage précède immédiatement celui de l'élément :current	4
E:future	Sélectionne un élément E dont l'affichage suit immédiatement celui de l'élément :current	4
E:drop	Sélectionne un élément E qui peut potentiellement recevoir un objet HTML dans le cadre d'un glissé-déposé (drag and drop)	4
E:drop(active)	Version alternative de E:drop. Sélectionne l'élément E qui est actuellement la cible d'un drag and drop pour l'élément que l'on est en train de faire glisser	4
E:drop(valid)	Version alternative de E:drop. Sélectionne l'élément E qui est actuellement la cible d'un drag and drop pour l'élément que l'on est en train de faire glisser uniquement si cet élément peut recevoir l'objet qui va y être déposé	4
E:drop(invalid)	Sélectionne tout élément E qui ne peut pas recevoir l'objet actuellement glissé mais qui pourrait recevoir un autre type d'objet	4
E:indeterminate	Sélectionne tout élément E de type input qui est dans un état indéterminé, c'est-à-dire ni coché ni non-coché	4
E:user-error	Sélectionne tout élément E de type input altéré par l'utilisateur et dont la valeur passée est invalide	4
E:blank	Sélectionne tout élément E ne possédant pas de contenu à part éventuellement des espaces	4
E:nth-column(n)	Sélectionne tout élément E représentant une cellule appartenant à la n-ième colonne d'un tableau ou d'une grille	4
E:nth-last-column(n)	Sélectionne tout élément E représentant une cellule appartenant à la n-ième colonne d'un tableau ou d'une grille en comptant à partir de la fin	4
:playing	Sélectionne tout média capable d'être joué (audio, vidéo…) lorsque l'élément en question est en train d'être joué (ou lu)	4
:paused	Sélectionne tout média capable d'être joué (audio, vidéo…) lorsque l'élément en question n'est pas en cours de lecture (en pause, en cours de chargement etc.)	4
E:matches(E1, .c1)	Sélectionne tout élément E qui est soit de type E1 ou qui possède un attribut class= »c1"	4
E:has(E1, .c1)	Sélectionne un élément E si au moins un des sélecteurs passés en paramètre correspond à l'élément (selon la portée, :scope, de l'élément)	4
E:dir(ltr)	Sélectionne tout élément E dont on à spécifié une direction de lecture « de gauche à droite » (la direction est à priori déterminée par la langue déclarée dans le document. Par exemple, pour le	4

Et le scraping dans tout ça ???

Comment des pratiques situées dans un espace virtuel peuvent-elles conduire les individus qui s'y identifient à un engagement social, culturel ou politique dans la société ?

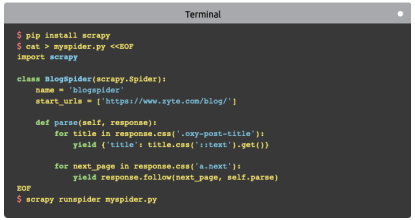
Comment rendre compte de la dynamique structurelle, informationnelle et cognitive à l'œuvre dans des groupes électroniques, notamment au sein des forums de discussion ?

Comment capter des masses de données colossales et hétérogènes ?



Ecosystème binaire des logiciels de scraping

Gratuits /

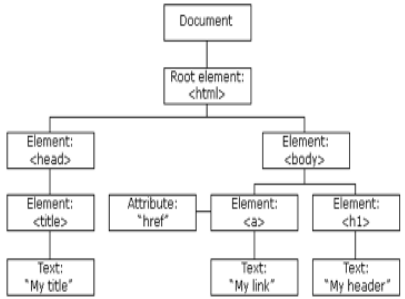


€€€ /

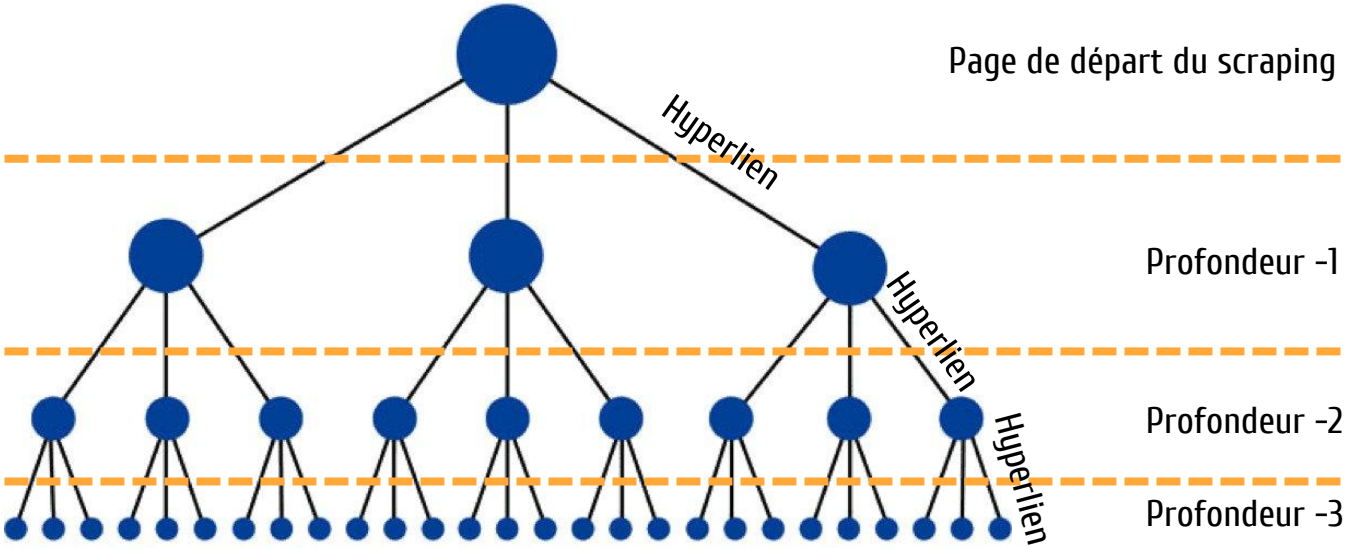
2 exemples de données à extraire :

- Articles sur un site de vente en ligne
- Messages sur un forum de discussion

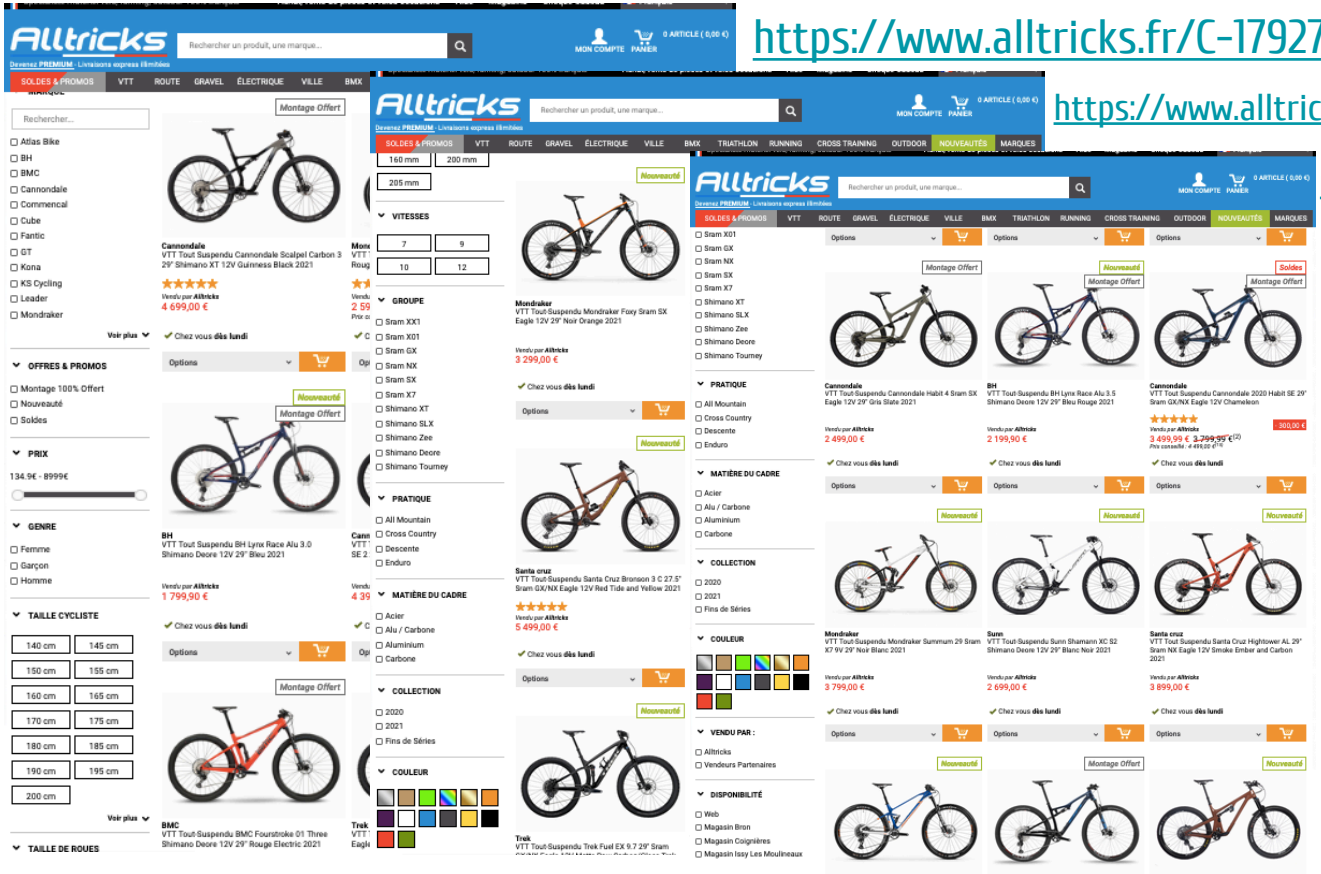
Le DOM



Arborescence des 2 sites web :



Un exemple simple : des VTT sur un site de vente en ligne

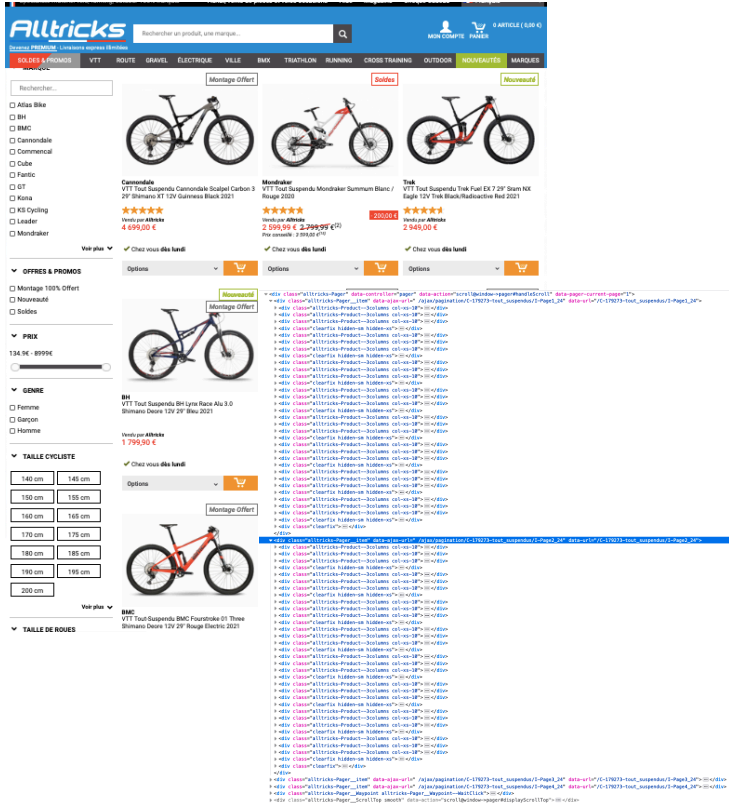


https://www.alltricks.fr/C-179273-tout_suspendus

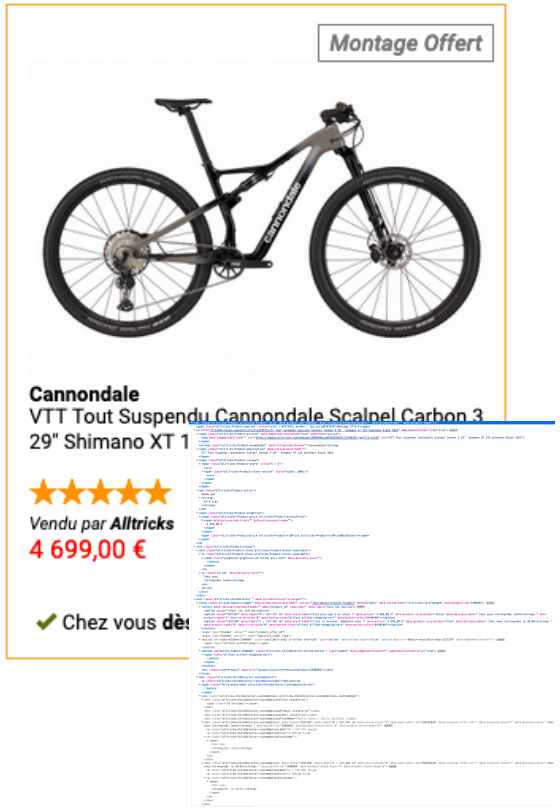
https://www.alltricks.fr/C-179273-tout_suspendus

https://www.alltricks.fr/C-179273-tout_suspendus

Un exemple simple : des VTT sur un site de vente en ligne

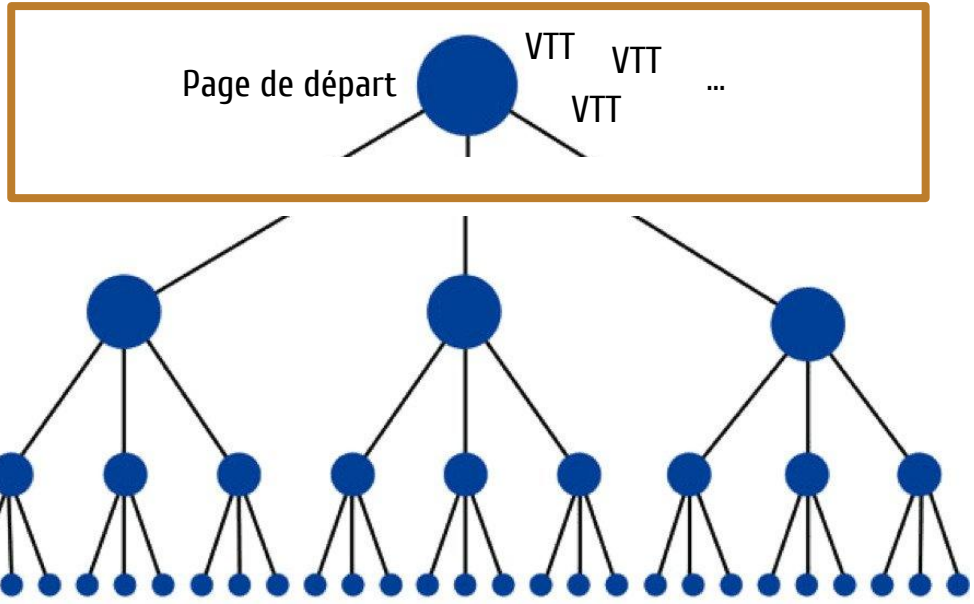


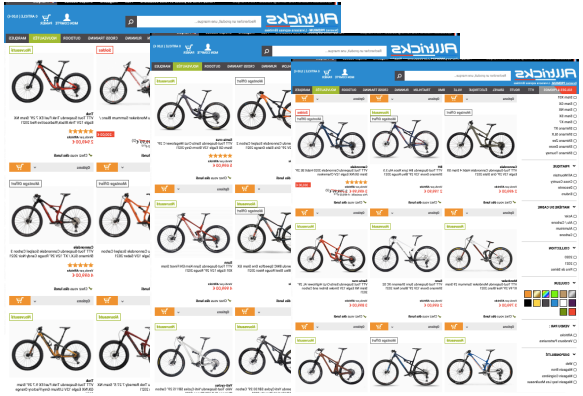
=



X 1000...

Un exemple simple : des VTT sur un site de vente en ligne





Un exemple complexe : des messages sur un forum de discussion

Un exemple complexe : des messages sur un forum de discussion

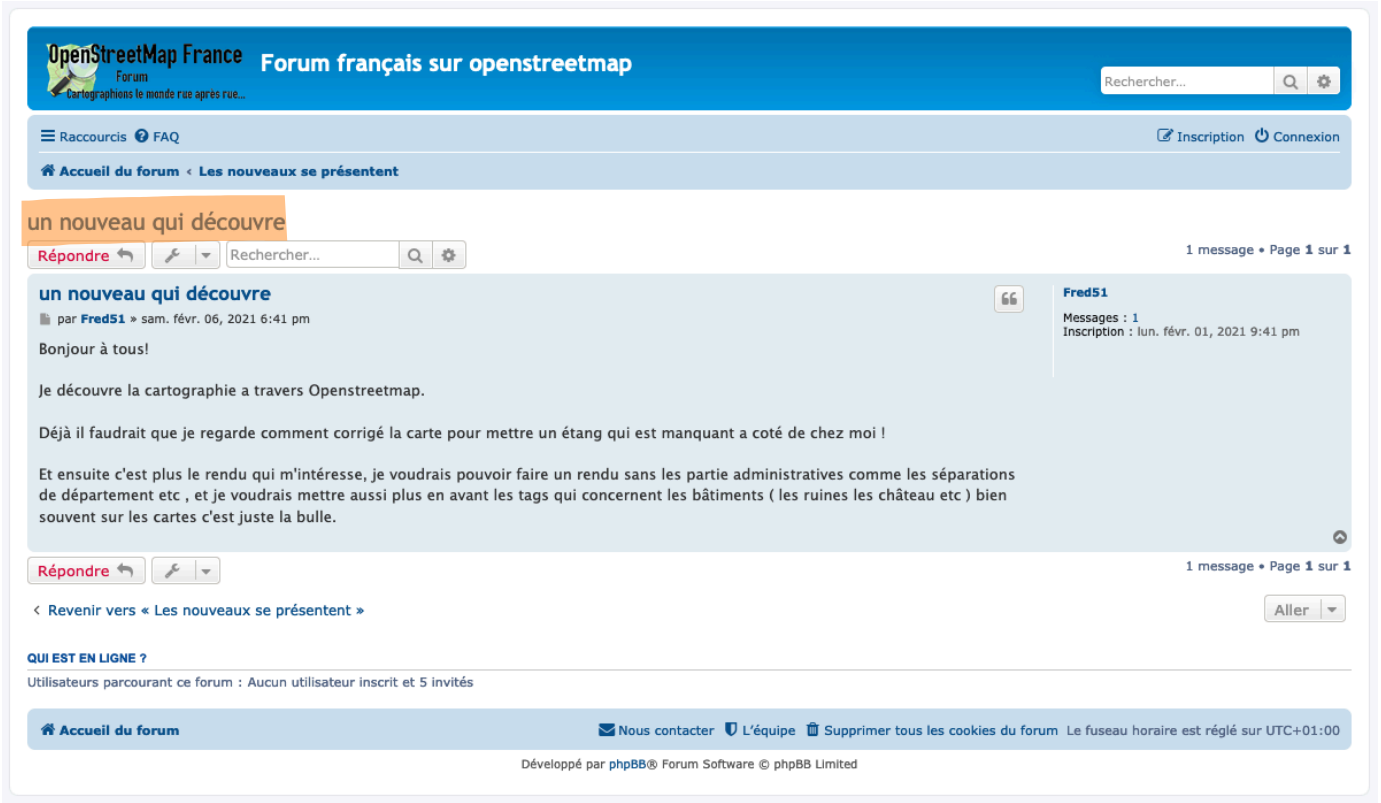
https://forum.openstreetmap.fr/

Un exemple complexe : des messages sur un forum de discussion

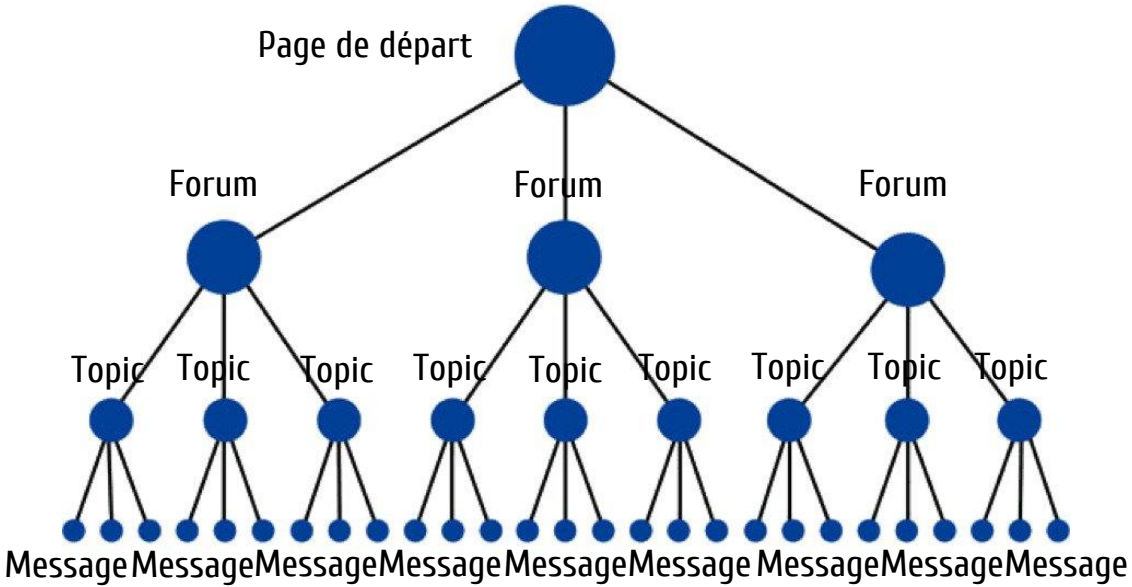
https://forum.openstreetmap.fr/viewforum.php?f=10

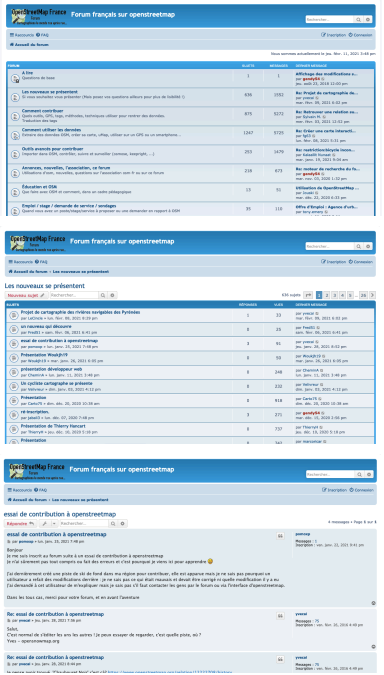
Un exemple complexe : des messages sur un forum de discussion

https://forum.openstreetmap.fr/viewforum.php?f=10&t=7534

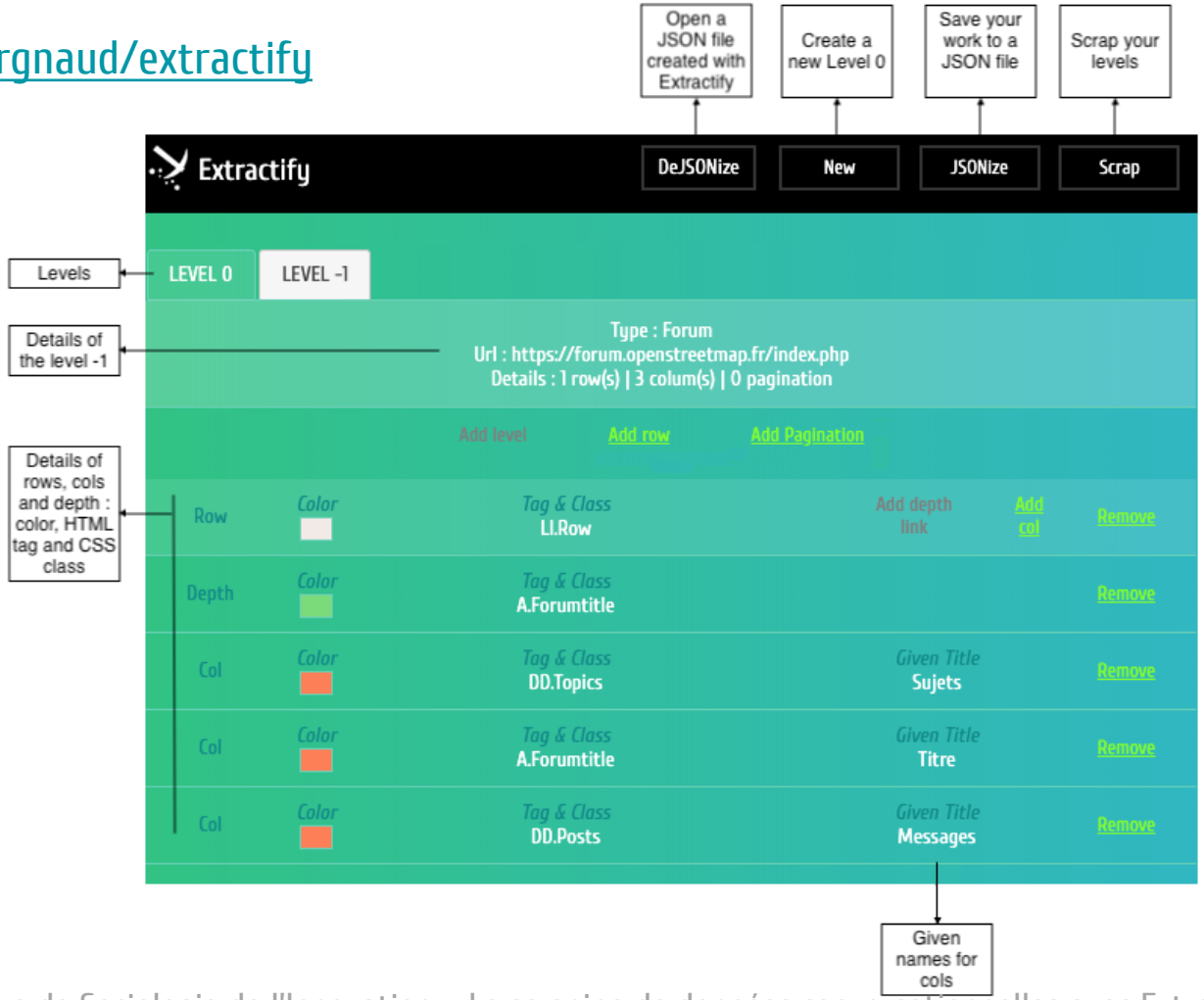
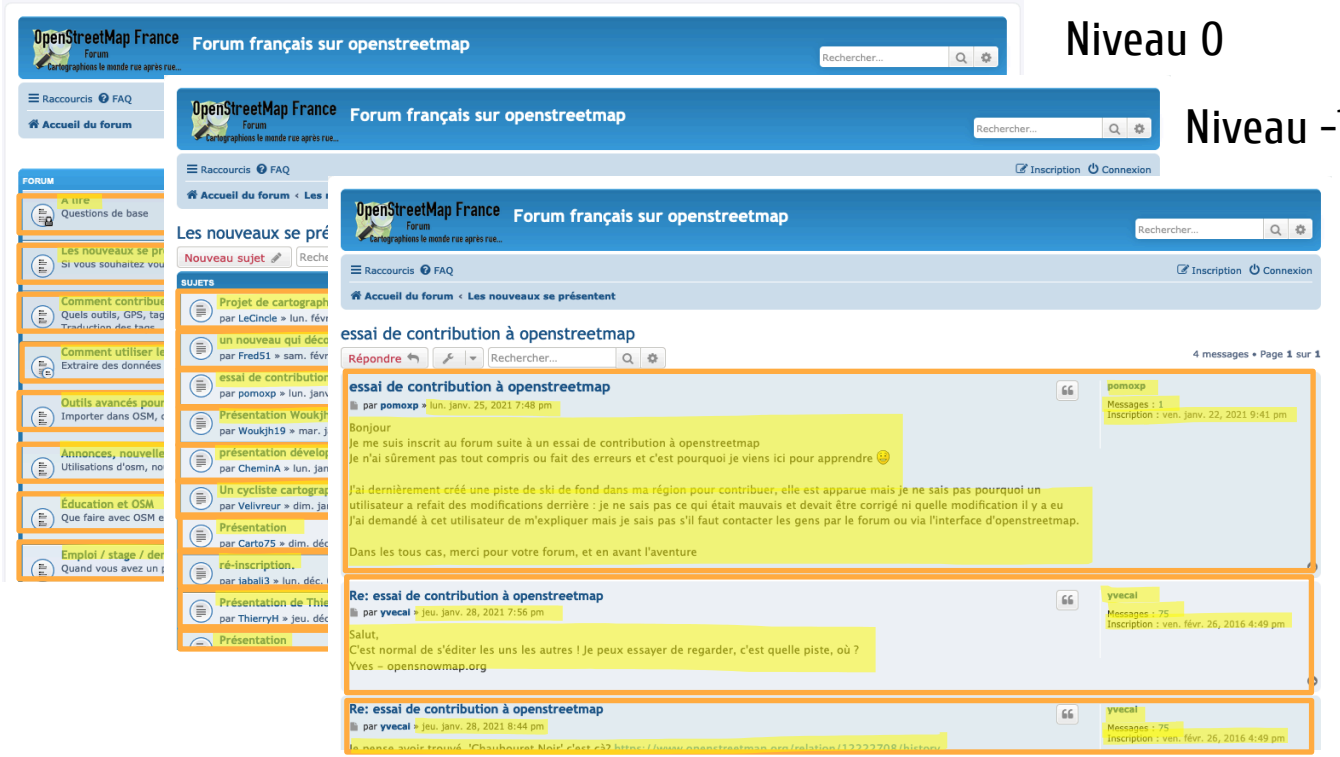


Un exemple complexe : des messages sur un forum de discussion





github.com/fredericvergnaud/extractify



github.com/fredericvergnaud/extractify