



Le scraping de données structurées sur le web à l'aide d'Extractify. Focus sur les données conversationnelles.

Frédéric Vergnaud

► To cite this version:

Frédéric Vergnaud. Le scraping de données structurées sur le web à l'aide d'Extractify. Focus sur les données conversationnelles.. Ecole Thématique “ Explo-SHS ”, Oct 2020, La Rochelle, France. hal-03052666

HAL Id: hal-03052666

<https://hal.science/hal-03052666>

Submitted on 10 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Ecole thématique 2020

EXPLO-SHS

12-16 oct. – La Rochelle



Le scraping de données structurées sur le web à l'aide d'Extractify.

Focus sur les données conversationnelles.

Frédéric Vergnaud

CSI – Centre de Sociologie de l'Innovation, i3 UMR CNRS 9217
Mines ParisTech, PSL Research University
[frederic.vergnaud\(at\)mines-paristech.fr](mailto:frederic.vergnaud@mines-paristech.fr)

École Thématique « EXPLO-SHS »

12-16 Octobre 2020

La Rochelle, France

Présentation

Extractify est une extension libre pour chrome, dont le but est de scraper des données structurées sur le web. Ce « plugin » est particulièrement conçu pour la récolte de commentaires ou de conversations en ligne tels que les forums.

Il permet de :

- 1) Sélectionner sur une page web des informations structurées sous forme de tableau (lignes x colonnes) de manière automatique, par sélection directe sur la page web, ou manuelle, en utilisant les sélecteurs CSS simples ou combinés.
- 2) Sélectionner la pagination de pages de structure identique et de même niveau
- 3) Recommencer le processus autant de fois que l'on veut pour des niveaux inférieurs
- 4) Scraper l'ensemble de la sélection
- 5) Obtenir finalement un fichier au format [json](#) facilement exportable vers d'autres logiciels

Installation manuelle pour Chrome

- Rendez-vous ici : <https://github.com/fredericvergnaud/extractify>
- Allez dans [Releases](#), téléchargez la dernière version, puis décompressez-là dans un répertoire de votre choix
- Ouvrez le navigateur Chrome (ou Chromium) et allez sur la page des extensions : `chrome://extensions/`
- Sélectionnez « Mode développeur » en haut à droite de la page
- Cliquez sur « Chargez l'extension non emballée » et sélectionnez le répertoire « extractify-numero_version » précédemment décompressé.

Utilisation

1) Introduction

D'un point de vue technique, Extractify utilise l'identification d'éléments HTML au sein de la structure d'une page web pour pouvoir en extraire les données. Pour repérer ces éléments, Extractify s'appuie sur les [sélecteurs CSS](#) qui utilisent certains des attributs des éléments HTML, principalement la classe et l'identifiant, afin de les repérer facilement dans la structure de la page web. Nous en verrons le détail un peu plus avant dans la présentation.

D'un point de vue conceptuel, pour Extractify, le niveau (level) est le degré de profondeur d'une page. La page sur laquelle vous allez exécuter pour la première fois le plugin constituera le niveau 0. Les pages plus profondes atteignables par hyperliens seront donc aux niveaux -1, puis -2, puis -3 ...etc.

Sur chaque page, Extractify considère que les données que vous voulez extraire sont structurées en lignes et en colonnes, c'est à dire sous la forme d'un tableau. Vous allez donc devoir d'abord sélectionner des lignes, puis des colonnes à l'intérieur de ces lignes.

Ensuite, vous pourrez sélectionner des liens de pagination, c'est à dire des liens vers des pages de profondeur équivalente à la page observée et structurées de la même manière que la page observée.

Enfin, vous pourrez ajouter un niveau inférieur en sélectionnant des liens vers ce niveau inférieur qui seront situés à l'intérieur des lignes déjà sélectionnées.

Extractify a été développé initialement pour extraire des données conversationnelles de forums en ligne.

En prenant un tel exemple typique de son utilisation, vous allez commencer par exécuter Extractify sur la page qui affichera l'ensemble des forums à scraper sous forme de liens. Cette page sera considérée comme le niveau 0.

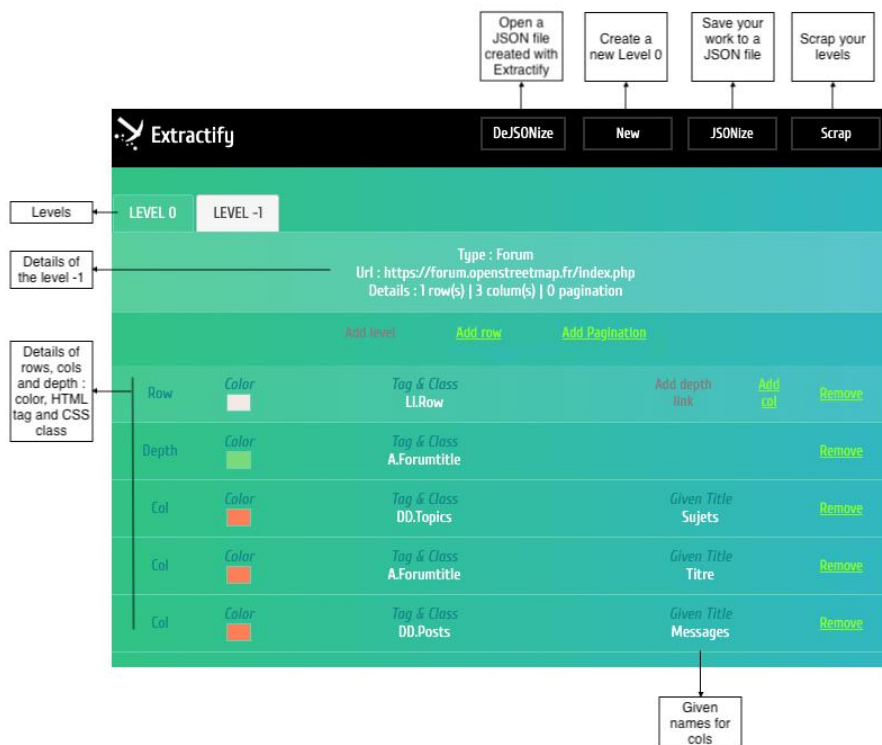
Chaque lien de cette page, menant à chaque forum, sera un lien vers un niveau inférieur (le niveau -1) de ce forum. Ce niveau inférieur affichera les topics.

Puis chaque lien menant à chaque topic sera lui-même un lien vers un niveau inférieur (le niveau -2) de ce topic. Ce niveau inférieur affichera les messages.

Ainsi, scraper une page web affichant des liens vers des forums reviendra à scraper 3 niveaux de profondeur :

- 1) Niveau 0 : la page affichant les forums, de laquelle, outre le lien du forum, on pourra extraire le titre du forum, son nombre de vues, son nombre de réponses ...etc.
- 2) Niveau -1 : les pages affichant les topics de ces forums, desquelles, outre les liens de topics, on pourra extraire le titre du topic, son nombre de vues, son nombre de réponses ...etc.
- 3) Niveau -2 : les pages affichant les messages de ces topics, desquelles on pourra extraire le nom de l'auteur, la date du message, le message ...etc.

2) Présentation de l'interface



3) Exemple de scraping d'un forum OpenStreetMap

Nous allons procéder à notre premier scraping ensemble sur le forum OpenStreetMap, qui est un forum de discussion en français autour d'OpenStreetMap, qui est un projet de carte ouverte et collaborative du monde.

Ce forum va nous permettre d'utiliser à la fois le mode automatique d'Extractify, mais également le mode manuel, en utilisant les sélecteurs CSS pour cibler certains champs que l'on voudrait récupérer.

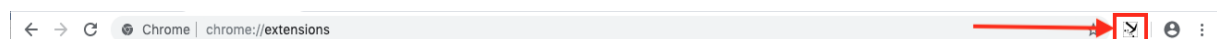
Vous pourrez ensuite vous exercer sur d'autres types de données à scraper.

1. **Rendez-vous** à l'adresse <https://forum.openstreetmap.fr/>

2. **Sélectionner un forum**

Le forum « Outils avancés pour contribuer » est un bon exemple car il regroupe 250 sujets pour un peu plus de 1450 messages, ce qui va être relativement rapide à scraper.

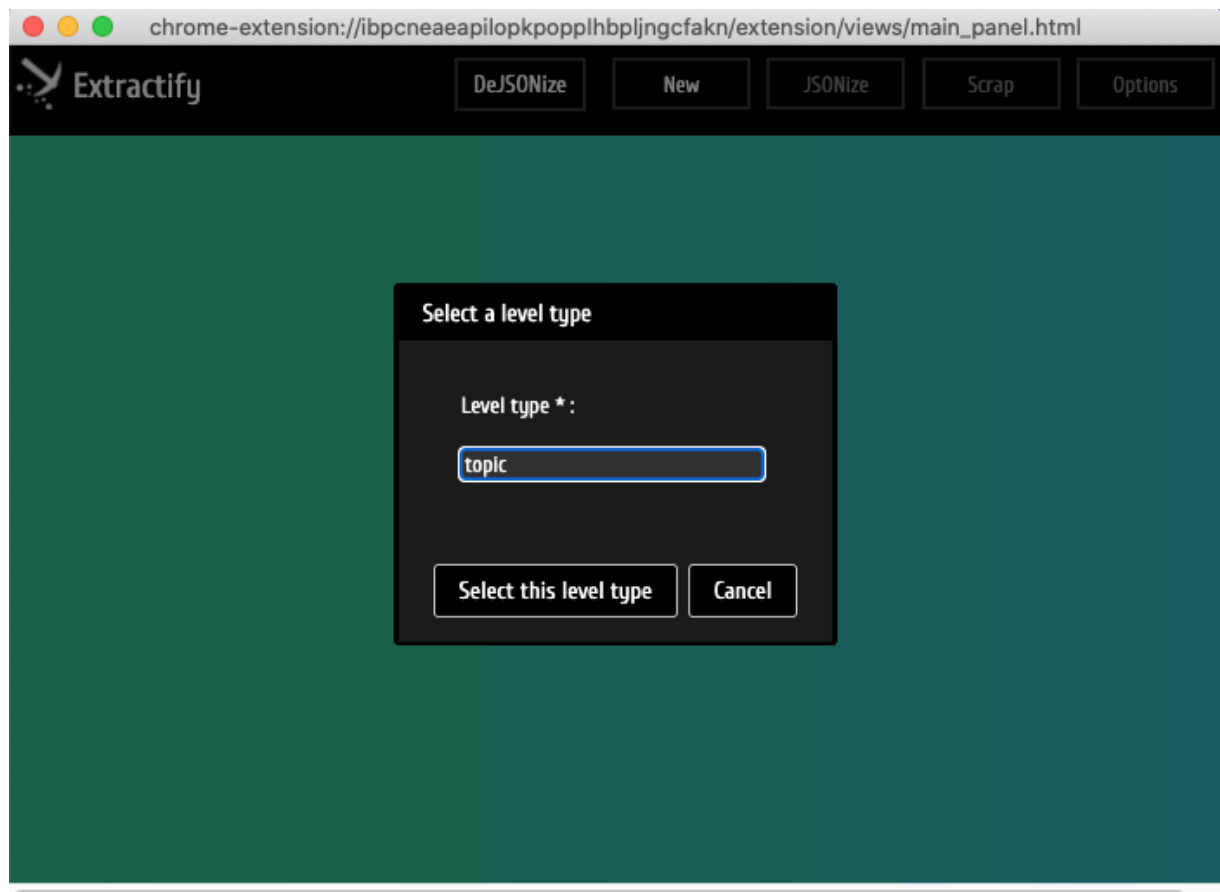
3. Cliquez sur l'icône du plugin dans la barre des extensions de chrome :



Le plugin s'ouvre sur le dialogue d'ajout du niveau 0.

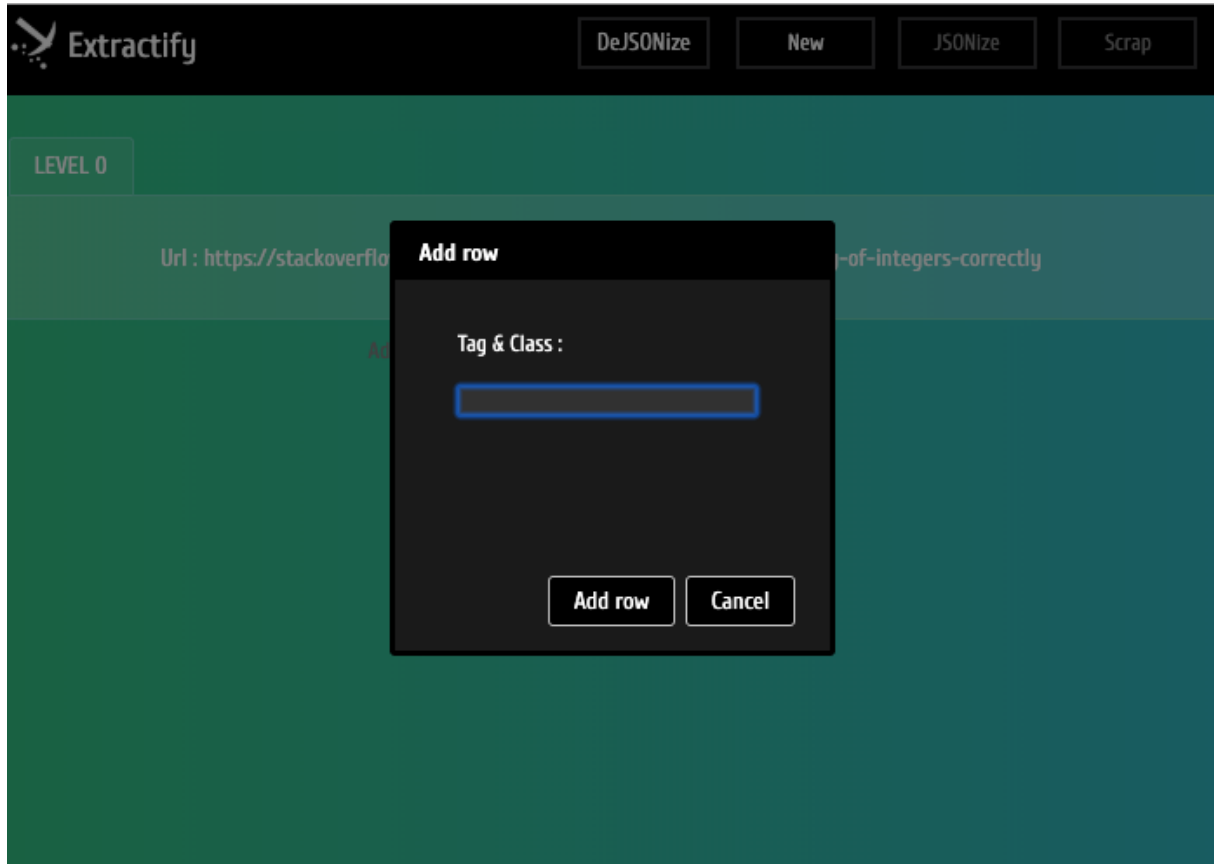
4. Ajout d'un niveau

Saisissez un type de niveau (obligatoire) puis cliquez sur « Select this level type ». Pour cet exemple nous allons saisir « topic » comme type de niveau :



5. Ajout de lignes

A l'ouverture du dialogue, cliquez directement sur le bouton « Add row » pour sélectionner les lignes **automatiquement**.



Déplacez ensuite votre souris sur la page web de votre navigateur et sélectionnez une ligne en cliquant une zone mise en surbrillance au survol de la structure HTML de la page :

Forum

Cartographiez le monde rue après rue...

OpenStreetMap France

Forum français sur openstreetmap

Rechercher...

Raccourcis

FAQ

Inscription

Connexion

Accueil du forum

< Outils avancés pour contribuer

Outils avancés pour contribuer

Nouveau sujet

Rechercher...

250 sujets

1

2

3

4

5

...

10

>

ANNONCES	RÉPONSES	VUES	DERNIER MESSAGE
<div><div></div><div>Projet du mois: les défibrillateurs</div><div>par gandy54 » ven. août 28, 2020 10:44 am » dans Comment contribuer</div></div>	0	1167	par gandy54 » ven. août 28, 2020 10:44 am

SUJETS	RÉPONSES	VUES	DERNIER MESSAGE
<div><div></div><div>Filtres Overpass Turbo et JOSM</div><div>par didwg » sam. août 29, 2020 9:34 pm</div></div>	0	871	par didwg » sam. août 29, 2020 9:34 pm
<div><div></div><div>Bâtiment complexe</div><div>par didwg » mar. août 04, 2020 7:24 pm</div></div>	0	0	par didwg » mar. août 04, 2020 7:24 pm
<div><div></div><div>Bâtiment complexe</div><div>par didwg » mar. août 04, 2020 7:17 pm</div></div>	0	0	par didwg » mar. août 04, 2020 7:17 pm
<div><div></div><div>Lignes superposées</div><div>par didwg » sam. août 01, 2020 8:32 pm</div></div>	2	337	par didwg » mar. août 04, 2020 7:12 pm
<div><div></div><div>[JOSM] Noeuds dupliqués sur des bâtiments</div><div>par MatthieuLyon69 » lun. août 03, 2020 2:47 pm</div></div>	2	411	par MatthieuLyon69 » mar. juin 27, 2020 3:46 pm
<div><div></div><div>JOSM - Couche IGN photographie 1950-1965 ?</div><div>par Jim005 » lun. juil. 27, 2020 12:45 am</div></div>	0	826	par Jim005 » lun. juil. 27, 2020 12:45 am
<div><div></div><div>Mise en œuvre via l'API d'OpenStreetMap pour permettre la géolocalisation</div><div>par Ubaxo21 » sam. avr. 20, 2019 1:12 pm</div></div>	1	2258	par alexosm » sam. juin 27, 2020 2:28 pm
<div><div></div><div>orthophoto en opendata</div><div>par tony.emery » mar. juin 16, 2020 10:00 am</div></div>	1	506	par cquest » mer. juin 17, 2020 7:33 am
<div><div></div><div>Regroupement de relations contigües</div><div>par gsimon » dim. juin 07, 2020 11:41 am</div></div>	7	786	par Romain » dim. juin 14, 2020 11:03 pm
<div><div></div><div>Comment fusionner ces trois requêtes OverPass ?</div><div>par percherie » mer. juin 03, 2020 1:50 pm</div></div>	2	827	par percherie » dim. juin 07, 2020 9:30 pm
<div><div></div><div>TMS Cadastre pour JOSM ne fonctionne plus</div><div>par percherie » ven. mai 29, 2020 8:49 am</div></div>	3	738	par Romain » sam. mai 30, 2020 9:30 pm
<div><div></div><div>Test accès BD Ortho IGN...</div><div>par cquest » ven. mai 27, 2016 6:10 pm</div></div>	63	46176	par wadouk » ven. mai 29, 2020 5:23 pm
<div><div></div><div>Comment retrouver un changeset sur une zone très limitée ?</div><div>par percherie » sam. mai 02, 2020 3:00 pm</div></div>	1	1083	par percherie » ven. mai 29, 2020 8:49 am
<div><div></div><div>Mise à jour Bano</div><div>par Captain47 » sam. mai 19, 2018 8:43 am</div></div>	10	5063	par cquest » ven. mai 29, 2020 8:37 am
<div><div></div><div>JOSM / téléchargement du cadastre vectorisé HS ?</div><div>par Maférik49 » mar. mai 19, 2020 10:06 pm</div></div>	1	425	par Maférik49 » mer. mai 20, 2020 5:49 pm
<div><div></div><div>Export des favoris Google Maps vers OpenStreetMap</div><div>par Géhéf » mar. mars 17, 2020 6:03 pm</div></div>	3	1043	par garenkreiz » ven. avr. 03, 2020 5:21 pm
<div><div></div><div>Comment positionner précisément un point isolé dont on connaît les coordonnées ?</div><div>par BTH » mer. janv. 29, 2020 11:36 am</div></div>	2	1205	par OsmO » jeu. janv. 30, 2020 10:41 am
<div><div></div><div>Cadastre et génération de fichier .OSM est HS ?</div><div>par Xavb22440 » ven. nov. 15, 2019 8:32 pm</div></div>	1	1548	par deuzeffe » sam. déc. 28, 2019 4:41 pm
<div><div></div><div>problème import adresses cadastre</div><div>par arobi » lun. oct. 21, 2019 6:28 pm</div></div>	3	1881	par deuzeffe » sam. déc. 28, 2019 4:40 pm

Au clic, les lignes identiques, c'est à dire les lignes qui ont la même structure HTML et CSS, se mettent également en surbrillance :

Forum

Cartographions le monde rue après rue...

OpenStreetMap France

Forum français sur openstreetmap

Rechercher...

Raccourcis

FAQ

Inscription

Connexion

Accueil du forum

Outils avancés pour contribuer

Outils avancés pour contribuer

Nouveau sujet

Rechercher...

250 sujets

1

2

3

4

5

...

10

>


ANNONCES	RÉPONSES	VUES	DERNIER MESSAGE
<div><div></div><div>Projet du mois: les défibrillateurs</div><div>par gendy54 » ven. août 28, 2020 10:44 am » dans Comment contribuer</div></div>	0	1167	par gendy54 » ven. août 28, 2020 10:44 am

SUJETS	RÉPONSES	VUES	DERNIER MESSAGE
<div><div></div><div>Filtres Overpass Turbo et JOSM</div><div>par didwg » sam. août 29, 2020 9:34 pm</div></div>	0	871	par didwg » sam. août 29, 2020 9:34 pm
<div><div></div><div>Bâtiment complexe</div><div>par didwg » mar. août 04, 2020 7:24 pm</div></div>	0	0	par didwg » mar. août 04, 2020 7:24 pm
<div><div></div><div>Bâtiment complexe</div><div>par didwg » mar. août 04, 2020 7:17 pm</div></div>	0	0	par didwg » mar. août 04, 2020 7:17 pm
<div><div></div><div>Lignes superposées</div><div>par didwg » sam. août 01, 2020 8:32 pm</div></div>	2	337	par didwg » mar. août 04, 2020 7:12 pm
<div><div></div><div>[JOSM] Noeuds dupliqués sur des bâtiments</div><div>par MatthieuLyon69 » lun. août 03, 2020 2:47 pm</div></div>	2	411	par MatthieuLyon69 » mar. août 04, 2020 3:46 pm
<div><div></div><div>JOSM - Couche IGN photographie 1950-1965 ?</div><div>par Jim005 » lun. juil. 27, 2020 12:45 am</div></div>	0	826	par Jim005 » lun. juil. 27, 2020 12:45 am
<div><div></div><div>Mise en œuvre via l'API d'OpenStreetMap pour permettre la géolocalisation</div><div>par Ubaxo21 » sam. avr. 20, 2019 1:12 pm</div></div>	1	2258	par alexosm » sam. juin 27, 2020 2:28 pm
<div><div></div><div>orthophoto en opendata</div><div>par tony.emery » mar. juin 16, 2020 10:00 am</div></div>	1	506	par cquest » mer. juin 17, 2020 7:33 am
<div><div></div><div>Regroupement de relations contigües</div><div>par gsimon » dim. juin 07, 2020 11:41 am</div></div>	7	786	par Romain » dim. juin 14, 2020 11:03 pm
<div><div></div><div>Comment fusionner ces trois requêtes OverPass ?</div><div>par percherie » mer. juin 03, 2020 1:50 pm</div></div>	2	827	par percherie » dim. juin 07, 2020 9:30 pm
<div><div></div><div>TMS Cadastre pour JOSM ne fonctionne plus</div><div>par percherie » ven. mai 29, 2020 8:49 am</div></div>	3	738	par Romain » sam. mai 30, 2020 9:30 pm
<div><div></div><div>Test accès BD Ortho IGN...</div><div>par cquest » ven. mai 27, 2016 6:10 pm</div></div>	63	46176	par wadouk » ven. mai 29, 2020 5:23 pm
<div><div></div><div>Comment retrouver un changeset sur une zone très limitée ?</div><div>par percherie » sam. mai 02, 2020 3:00 pm</div></div>	1	1083	par percherie » ven. mai 29, 2020 8:49 am
<div><div></div><div>Mise à jour Bano</div><div>par Captain47 » sam. mai 19, 2018 8:43 am</div></div>	10	5063	par cquest » ven. mai 29, 2020 8:37 am
<div><div></div><div>JOSM / téléchargement du cadastre vectorisé HS ?</div><div>par Maferik49 » mar. mai 19, 2020 10:06 pm</div></div>	1	425	par Maferik49 » mer. mai 20, 2020 5:49 pm
<div><div></div><div>Export des favoris Google Maps vers OpenStreetMap</div><div>par Géhéf » mar. mars 17, 2020 6:03 pm</div></div>	3	1043	par garenkreiz » ven. avr. 03, 2020 5:21 pm
<div><div></div><div>Comment positionner précisément un point isolé dont on connaît les coordonnées ?</div><div>par BTH » mer. janv. 29, 2020 11:36 am</div></div>	2	1205	par OsmO » jeu. janv. 30, 2020 10:41 am

On remarque ici qu'une ligne sur deux a été sélectionnée. Regardons pourquoi à l'aide de l'inspecteur de code de Chrome. Cela nous donne par ailleurs une bonne occasion d'utiliser cet outil qui se révèle indispensable lorsque l'on veut scraper des données web.

Chaque portion de texte sur une page web est encapsulée dans des balises HTML. Ces balises possèdent des attributs qui vont nous permettre de les cibler plus facilement afin d'en extraire le contenu. C'est à l'aide de l'inspecteur interne de Chrome que nous allons pouvoir détailler la structure HTML de la page web, et ainsi repérer les balises et classes à cibler.

Faites un clic-droit sur une ligne non sélectionnée puis choisissez « Inspectez » :



OpenStreetMap France

Forum

Cartographier le monde rue après rue...

Forum français sur openstreetmap

Rechercher...



Raccourcis
FAQ
Inscription
Connexion





































Accueil du forum
Outils avancés pour contribuer


Outils avancés pour contribuer

Nouveau sujet
Rechercher...

250 sujets
1 2 3 4 5 ... 10

ANNONCES		RÉPONSES	VUES	DERNIER MESSAGE
 Projet du mois: les défilibrateurs par gendy54 » ven. août 28, 2020 10:44 am » dans Comment contribuer		0	1167	par gendy54  ven. août 28, 2020 10:44 am

SUJETS		RÉPONSES	VUES	DERNIER MESSAGE
 Filtres Overpass Turbo et JOSM par didwg » sam. août 29, 2020 9:34 pm		0	871	par didwg  sam. août 29, 2020 9:34 pm
 Bâtiment complexe par didwg » mar. août 04, 2020 7:24 pm		0	0	par didwg  mar. août 04, 2020 7:24 pm
 Bâtiment complexe par didwg » mar. août 04, 2020 7:17 pm		0	0	par didwg  mar. août 04, 2020 7:17 pm
 Lignes superposées par didwg » sam. août 01, 2020 8:32 pm		2	337	par didwg  mar. août 04, 2020 7:12 pm
 [JOSM] Nœuds dupliqués sur des bâtiments  par MatthieuLyon69 » lun. août 03, 2020 2:47 pm		2	411	par MatthieuLyon69  mar. août 04, 2020 3:46 pm
 JOSM - Couche IGN photographie 1950-1965 ? par Jim005 » lun. juil. 27, 2020 12:45 am		0	826	par Jim005  lun. juil. 27, 2020 12:45 am
 Mise en œuvre via l'API d'OpenStreetMap pour permettre la géolocalisation par Ubaxo21 » sam. avr. 20, 2019 1:12 pm		1	2258	par alexosm  sam. juin 27, 2020 2:28 pm
 orthophoto en opendata par tony.emery » mar. juin 16, 2020 10:00 am			506	par cquest  mer. juin 17, 2020 7:33 am
 Regroupement de relations contiguës par gsimon » dim. juin 07, 2020 11:41 am			786	par Romain  dim. juin 14, 2020 11:03 pm
 Comment fusionner ces trois requêtes OverPass ? par percherie » mer. juin 03, 2020 1:50 pm			827	par percherie  dim. juin 07, 2020 9:30 pm
 TMS Cadastre pour JOSM ne fonctionne plus par percherie » ven. mai 29, 2020 8:49 am			738	par Romain  sam. mai 30, 2020 9:30 pm
 Test accès BD Ortho IGN... par cquest » ven. mai 27, 2016 6:10 pm			46176	par wadoud  ven. mai 29, 2020 5:23 pm
 Comment retrouver un changement sur une zone très limitée par percherie » sam. mai 02, 2020 3:00 pm			1083	par percherie  ven. mai 29, 2020 8:49 am
 Mise à jour Bano par Captain47 » sam. mai 19, 2018 8:43 am			5063	par cquest  ven. mai 29, 2020 8:37 am
 JOSM / téléchargement du cadastre vectorisé HS ? par Maferik49 » mar. mai 19, 2020 10:06 pm		1	425	par Maferik49  mer. mai 20, 2020 5:49 pm
 Export des favoris Google Maps vers OpenStreetMap par Géhéf » mar. mars 17, 2020 6:03 pm		3	1043	par garenkreiz  ven. avr. 03, 2020 5:21 pm
 Comment positionner précisément un point isolé dont on connaît les coordonnées ? par BTH » mer. janv. 29, 2020 11:36 am		2	1205	par OsmO  jeu. janv. 30, 2020 10:41 am
 Cadastre et génération de fichier .OSM est HS ?				

Retour
Avancer
Actualiser
Enregistrer sous...
Imprimer...
Caster...
Traduire en français
 **AdBlock** — le meilleur bloqueur de pubs
Afficher le code source de la page
Inspector
Voix

L'inspecteur s'ouvre directement sur la portion HTML que vous venez de sélectionner :

Forum français sur openstreetmap
Cartographions le monde rue après rue...

Rechercher...

Raccourcis FAQ
Inscription Connexion

Accueil du forum Outils avancés pour contribuer

Outils avancés pour contribuer

Nouveau sujet Rechercher...
250 sujets
1 2 3 4 5 ... 10 >

ANNONCES	RÉPONSES	VUES	DERNIER MESSAGE
<i>i</i> Projet du mois: les défilibrateurs par gendy54 » ven. août 28, 2020 10:44 am » dans Comment contribuer	0	1167	par gendy54 ven. août 28, 2020 10:44 am

SUJETS	RÉPONSES	VUES	DERNIER MESSAGE
<i>i</i> Filtres Overpass Turbo et JOSM par didwg » sam. août 29, 2020 9:34 pm	0	871	par didwg sam. août 29, 2020 9:34 pm
<i>i</i> Bâtiment complexe par didwg » mar. août 04, 2020 7:24 pm	0	0	par didwg mar. août 04, 2020 7:24 pm
<i>i</i> Bâtiment complexe par didwg » mar. août 04, 2020 7:17 pm	0	0	par didwg mar. août 04, 2020 7:17 pm
<i>i</i> Lignes superposées par didwg » sam. août 01, 2020 8:32 pm	2	337	par didwg mar. août 04, 2020 7:12 pm
<i>i</i> [JOSM] Nœuds dupliqués sur des bâtiments par MatthieuLyon69 » lun. août 03, 2020 2:47 pm	2	411	par MatthieuLyon69 mar. août 04, 2020 3:46 pm
<i>i</i> JOSM - Couche IGN photographie 1950-1965 ? par Jim005 » lun. juil. 27, 2020 12:45 am	0	826	par Jim005 lun. juil. 27, 2020 12:45 am
<i>i</i> Mise en œuvre via l'API d'OpenStreetMap pour permettre la géolocalisation par Ubaxo21 » sam. avr. 20, 2019 1:12 pm	1	2258	par alexosm sam. juin 27, 2020 2:28 pm
<i>i</i> orthophoto en opendata par tony.emery » mar. juin 16, 2020 10:00 am	1	506	par cquest mer. juin 17, 2020 7:33 am
<i>i</i> Regroupement de relations contigües par gsimon » dim. juin 07, 2020 11:41 am	7	786	par Romain dim. juin 14, 2020 11:03 pm
<i>i</i> Comment fusionner ces trois requêtes OverPass ? par percherie » mer. juin 03, 2020 1:50 pm	2	827	par percherie dim. juin 07, 2020 9:30 pm

Elements Console Sources Network Performance Memory Application Security Lighthouse AdBlock

Styles Computed Event Listeners DOM Breakpoints Properties >>

Filter : #hov .cls +

```

element.style {
}

dl.row-item dt, dl.row-item dd {
  min-height: 35px;
}

dl.row-item dt {
  background-repeat: no-repeat;
  background-position: 5px 95%;
}

ul.topiclist dt {
  width: 100%;
  margin-right: -440px;
  font-size: 1.1em;
}

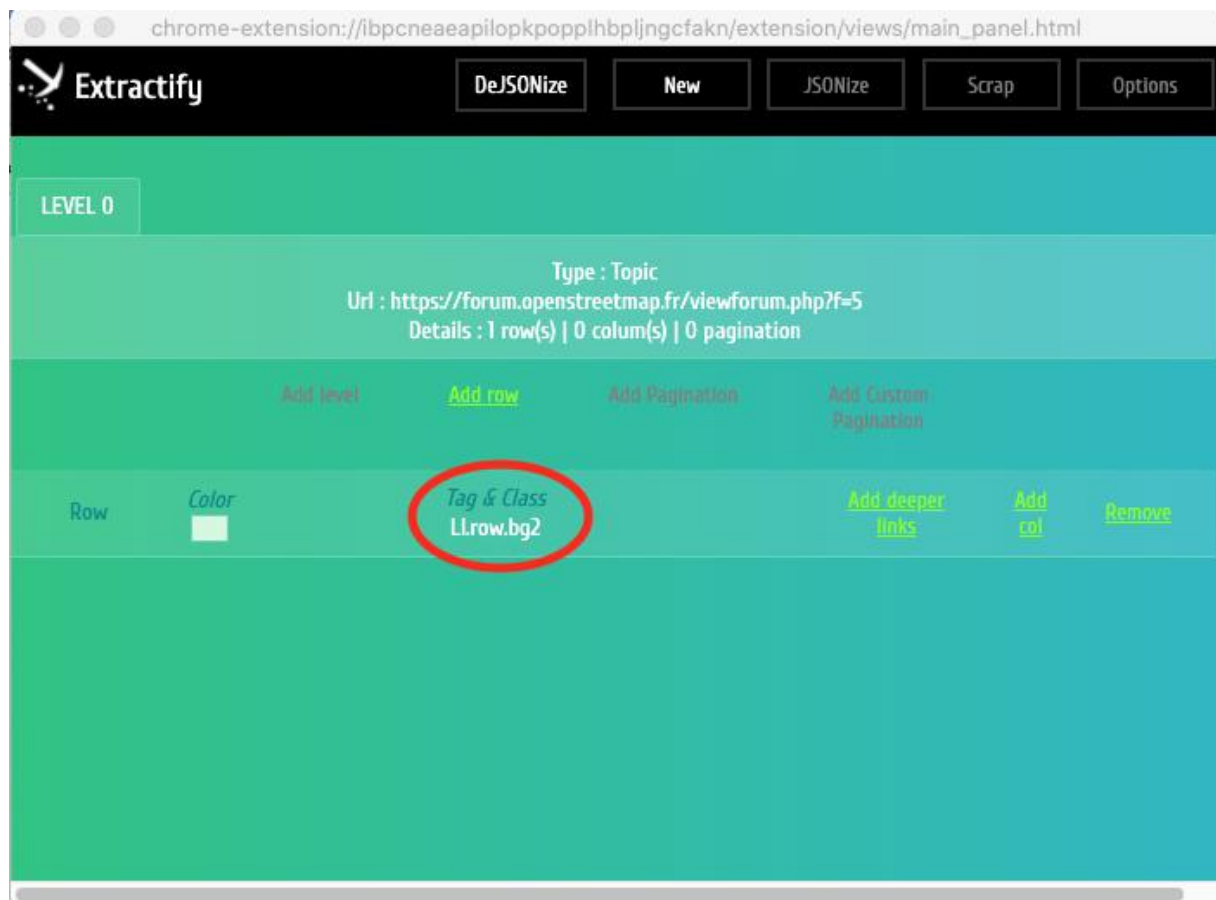
ul.topiclist dt, ul.topiclist dd {
  display: block;
  float: left;
}

dt {
  display: block;
}

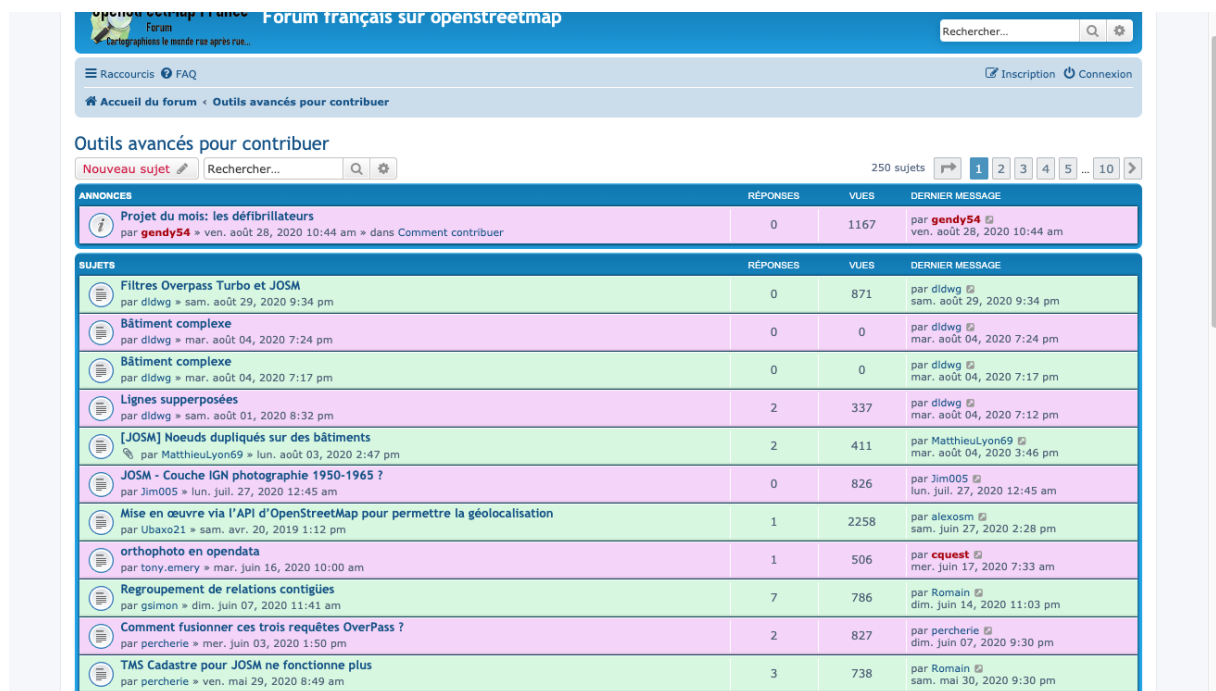
Inherited from li.row.bg1
ul.topiclist li {
  color: #4C5D77;
}
```

En survolant la structure HTML, on s'aperçoit que les lignes de topics sont structurées par des balises . Chaque balise possède un attribut « class ». Et c'est cet attribut qu'a ciblé Extractify pour sélectionner les lignes. En fonction de la ligne que vous avez survolée puis cliquée, vous avez sélectionné soit une balise de « class » « row » et « bg1 », soit une balise de « class » « row » et « bg2 ».

Et effectivement, on le visualise comme tel sur l'interface d'Extractify :



Si l'on veut la totalité des lignes, il faut donc recommencer l'opération d'ajout en sélectionnant les lignes « li » de classe « row bg2 » non sélectionnées :




Extractify

DeJSONize

New

JSONize

Scrap

Options

LEVEL 0

Type : Topic

Url : <https://forum.openstreetmap.fr/viewforum.php?f=5>


Details : 2 row(s) | 0 colum(s) | 0 pagination

Add level

Add row

Add Pagination

Add Custom
Pagination

Row	Color 	Tag & Class LI.row.bg2	Add deeper links	Add col	Remove
Row	Color 	Tag & Class LI.row.bg1	Add deeper links	Add col	Remove

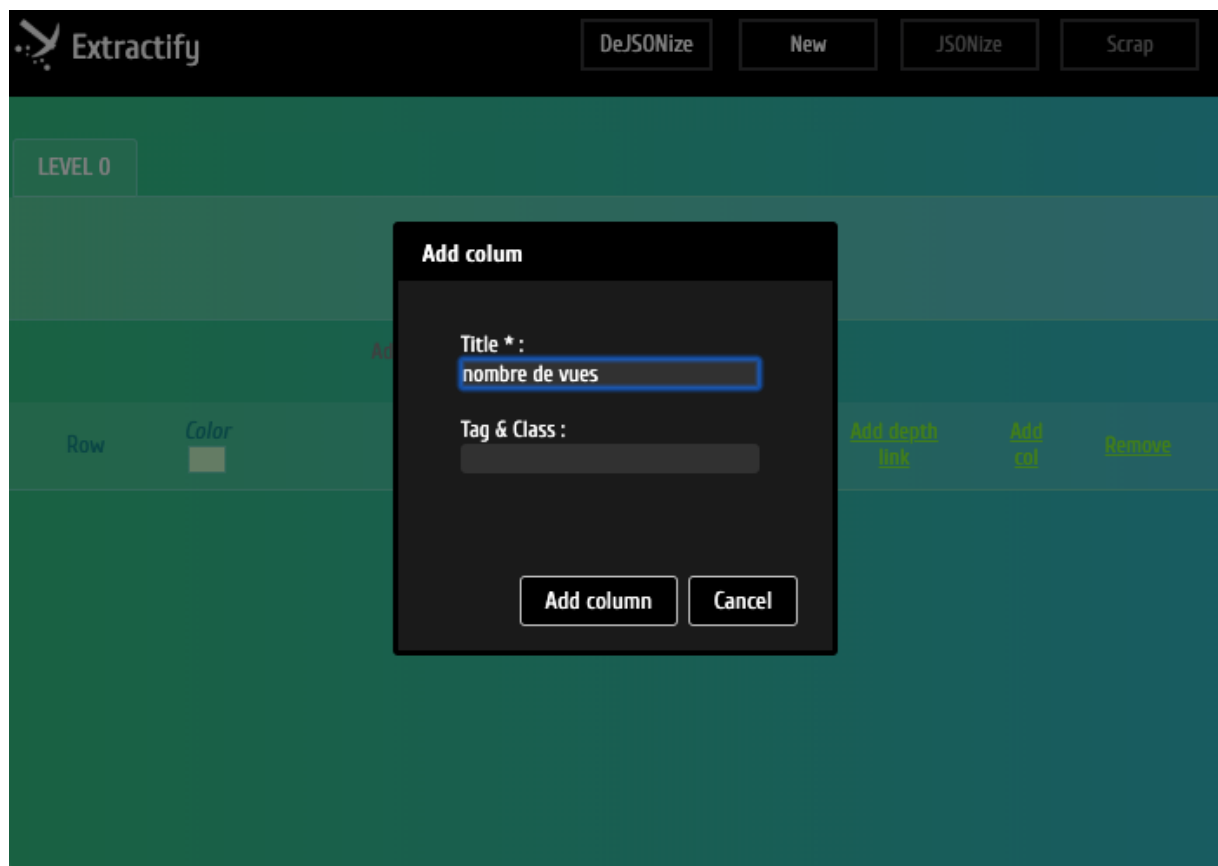
Il est à noter ici qu'une stratégie plus efficace et plus rapide aurait été de sélectionner directement tous les éléments « li » de classe « row ». Ainsi, la totalité des lignes aurait été sélectionnée d'un coup.

6. Ajout de colonnes

L'ajout de colonnes suit le même principe que l'ajout de lignes, à part le fait que vous devez saisir obligatoirement un titre de colonne.

Dans notre exemple, nous n'allons sélectionner que le titre du topic, ainsi que le nombre de vues. Le nombre de réponses étant de toute manière le nombre de messages du topic, ce que nous obtiendrons par le scraping. De la même manière, nous n'avons pas besoin des informations sur le dernier message, car nous retrouverons celui-ci dans le scraping futur de la totalité des messages !


Vous allez donc cliquer sur « Add col » de la première ligne :



Après avoir donner un titre à votre colonne, cliquez directement sur « Add column » de la boîte de dialogue pour exécuter la sélection en mode automatique, et sélectionner sur la page web la case « Vues ».

Comme vous avez 2 types de lignes, il va falloir répéter cette séquence pour l'autre type, car Extractify ne peut sélectionner que les colonnes qui sont situées à l'intérieur des lignes déjà ciblées.



Et vous répétez la sélection pour le titre de topic :



Forum
Cartographions le monde rue après rue...

Forum français sur openstreetmap

Rechercher...



Raccourcis

FAQ

Inscription

Connexion

Accueil du forum • Outils avancés pour contribuer

Outils avancés pour contribuer

Nouveau sujet

Rechercher...



250 sujets

1

2

3

4

5

...


10

>

ANNONCES	RÉPONSES	VUES	DERNIER MESSAGE
<div><div>Projet du mois: les défibrillateurs par gendy54 » ven. août 28, 2020 10:44 am » dans Comment contribuer</div></div>	0	1168	par gendy54  ven. août 28, 2020 10:44 am

SUJETS	RÉPONSES	VUES	DERNIER MESSAGE
<div><div>Filtres Overpass Turbo et JOSM par didwg » sam. août 29, 2020 9:34 pm</div></div>	0	872	par didwg  sam. août 29, 2020 9:34 pm
<div><div>Bâtiment complexe par didwg » mar. août 04, 2020 7:24 pm</div></div>	0	0	par didwg  mar. août 04, 2020 7:24 pm
<div><div>Bâtiment complexe par didwg » mar. août 04, 2020 7:17 pm</div></div>	0	0	par didwg  mar. août 04, 2020 7:17 pm
<div><div>Lignes superposées par didwg » sam. août 01, 2020 8:32 pm</div></div>	2	337	par didwg  mar. août 04, 2020 7:12 pm
<div><div>[JOSM] Noeuds dupliqués sur des bâtiments  par MatthieuLyon69 » lun. août 03, 2020 2:47 pm</div></div>	2	411	par MatthieuLyon69  mar. août 04, 2020 3:46 pm
<div><div>JOSM - Couche IGN photographie 1950-1965 ? par Jim005 » lun. juil. 27, 2020 12:45 am</div></div>	0	828	par Jim005  lun. juil. 27, 2020 12:45 am
<div><div>Mise en œuvre via l'API d'OpenStreetMap pour permettre la géolocalisation par Ubaxo21 » sam. avr. 20, 2019 1:12 pm</div></div>	1	2258	par alexosm  sam. juin 27, 2020 2:28 pm
<div><div>orthophoto en opendata par tony.emery » mar. juin 16, 2020 10:00 am</div></div>	1	506	par cquest  mer. juin 17, 2020 7:33 am
<div><div>Regroupement de relations contigües par gsimon » dim. juin 07, 2020 11:41 am</div></div>	7	786	par Romain  dim. juin 14, 2020 11:03 pm
<div><div>Comment fusionner ces trois requêtes OverPass ? par percherie » mer. juin 03, 2020 1:50 pm</div></div>	2	830	par percherie  dim. juin 07, 2020 9:30 pm
<div><div>TMS Cadastre pour JOSM ne fonctionne plus par percherie » ven. mai 29, 2020 8:49 am</div></div>	3	738	par Romain  sam. mai 30, 2020 9:30 pm
<div><div>Test accès BD Ortho IGN... par cquest » ven. mai 27, 2016 6:10 pm</div></div>	<div><div>1</div><div><div>2</div></div><div><div>3</div></div></div> <div>63</div>	46178	par wadouk  ven. mai 29, 2020 5:23 pm
<div><div>Comment retrouver un changeset sur une zone très limitée ? par percherie » sam. mai 02, 2020 3:00 pm</div></div>	1	1083	par percherie  ven. mai 29, 2020 8:49 am
<div><div>Mise à jour Band par Captain47 » sam. mai 19, 2018 8:43 am</div></div>	10	5064	par cquest  ven. mai 29, 2020 8:37 am
<div><div>JOSM / téléchargement du cadastre vectorisé HS ? par Maférik49 » mar. mai 19, 2020 10:06 pm</div></div>	1	427	par Maférik49  mer. mai 20, 2020 5:49 pm
<div><div>Export des favoris Google Maps vers OpenStreetMap par Géhéf » mar. mars 17, 2020 6:03 pm</div></div>	3	1045	par garenkreiz  ven. avr. 03, 2020 5:21 pm
<div><div>Comment positionner précisément un point isolé dont on connaît les coordonnées ? par BTH » mer. janv. 29, 2020 11:36 am</div></div>	2	1205	par OsmO  jeu. janv. 30, 2020 10:41 am
<div><div>Cadastre et génération de fichier .OSM est HS ? par Xavb22440 » ven. nov. 15, 2019 8:32 pm</div></div>	1	1550	par deuzeffe  sam. déc. 28, 2019 4:41 pm

chrome-extension://ibpcneaeapilopkpopplhbpljngcfakn/extension/views/main_panel.html

 Extractify

DeJSONize

New

JSONize

Scrap

Options

LEVEL 0

Type : Topic

Url : https://forum.openstreetmap.fr/viewforum.php?f=5

Details : 2 row(s) | 4 colum(s) | 0 pagination

Add level

Add row

Add Pagination

Add Custom
Pagination

Row	<div>Color</div> <div></div>	<div>Tag & Class</div> <div>LL.row.bg2</div>	<div>Add deeper links</div>	<div>Add col</div>	<div>Remove</div>
Col	<div>Color</div> <div></div>	<div>Tag & Class</div> <div>DD.views</div>	<div>Given Title</div> <div>Nbre de vues</div>		<div>Remove</div>
Col	<div>Color</div> <div></div>	<div>Tag & Class</div> <div>A.topictitle</div>	<div>Given Title</div> <div>Titre topic</div>		<div>Remove</div>
Row	<div>Color</div> <div></div>	<div>Tag & Class</div> <div>LL.row.bg1</div>	<div>Add deeper links</div>	<div>Add col</div>	<div>Remove</div>
Col	<div>Color</div> <div></div>	<div>Tag & Class</div> <div>DD.views</div>	<div>Given Title</div> <div>Nbre de vues</div>		<div>Remove</div>

7. Ajout d'une pagination

Si vous voulez scraper des pages de même niveau, mais situées à des adresses différentes accessibles *via* des liens de pagination, vous pouvez le faire en cliquant sur « Select Pagination » qui suit le même principe que l'ajout de lignes.

OpenStreetMap France

Forum

Cartographier le monde rue après rue...

Forum français sur openstreetmap

Rechercher...

Raccourcis

FAQ

Inscription

Connexion

Accueil du forum < Outils avancés pour contribuer

Outils avancés pour contribuer

Nouveau sujet

Rechercher...

250 sujets

1

2

3

4

5

...


10

ANNONCES	RÉPONSES	VUES	DERNIER MESSAGE
<div><div></div><div>Projet du mois: les défibrillateurs</div><div>par gendy54 » ven. août 28, 2020 10:44 am » dans Comment contribuer</div></div>	0	1188	par gendy54 ven. août 28, 2020 10:44 am

SUJETS	RÉPONSES	VUES	DERNIER MESSAGE
<div><div></div><div>Filtres Overpass Turbo et JOSM</div><div>par didwg » sam. août 29, 2020 9:34 pm</div></div>	0	876	par didwg sam. août 29, 2020 9:34 pm
<div><div></div><div>Bâtiment complexe</div><div>par didwg » mar. août 04, 2020 7:24 pm</div></div>	0	0	par didwg mar. août 04, 2020 7:24 pm
<div><div></div><div>Bâtiment complexe</div><div>par didwg » mar. août 04, 2020 7:17 pm</div></div>	0	0	par didwg mar. août 04, 2020 7:17 pm
<div><div></div><div>Lignes superposées</div><div>par didwg » sam. août 01, 2020 8:32 pm</div></div>	2	341	par didwg mar. août 04, 2020 7:12 pm
<div><div></div><div>[JOSM] Noeuds dupliqués sur des bâtiments</div><div>par MatthieuLyon69 » lun. août 03, 2020 2:47 pm</div></div>	2	421	par MatthieuLyon69 mar. août 04, 2020 3:46 pm
<div><div></div><div>JOSM - Couche IGN photographie 1950-1965 ?</div><div>par Jim005 » lun. juil. 27, 2020 12:45 am</div></div>	0	831	par Jim005 lun. juil. 27, 2020 12:45 am
<div><div></div><div>Mise en œuvre via l'API d'OpenStreetMap pour permettre la géolocalisation</div><div>par Ubaxo21 » sam. avr. 20, 2019 1:12 pm</div></div>	1	2262	par alexosm sam. juin 27, 2020 2:28 pm
<div><div></div><div>orthophoto en opendata</div><div>par tony.emery » mar. juin 16, 2020 10:00 am</div></div>	1	511	par cquest mer. juin 17, 2020 7:33 am
<div><div></div><div>Regroupement de relations contigües</div><div>par gsimon » dim. juin 07, 2020 11:41 am</div></div>	7	792	par Romain dim. juin 14, 2020 11:03 pm
<div><div></div><div>Comment fusionner ces trois requêtes OverPass ?</div><div>par percherie » mer. juin 03, 2020 1:50 pm</div></div>	2	833	par percherie dim. juin 07, 2020 9:30 pm
<div><div></div><div>TMS Cadastre pour JOSM ne fonctionne plus</div><div>par percherie » ven. mai 29, 2020 8:49 am</div></div>	3	741	par Romain sam. mai 30, 2020 9:30 pm
<div><div></div><div>Test accès BD Ortho IGN...</div><div>par cquest » ven. mai 27, 2016 6:10 pm</div></div>	63	46189	par wadouk ven. mai 29, 2020 5:23 pm
<div><div></div><div>Comment retrouver un changeset sur une zone très limitée ?</div><div>par percherie » sam. mai 02, 2020 3:00 pm</div></div>	1	1087	par percherie ven. mai 29, 2020 8:49 am
<div><div></div><div>Mise à jour Bang</div><div>par Captain47 » sam. mai 19, 2018 8:43 am</div></div>	10	5070	par cquest ven. mai 29, 2020 8:37 am
<div><div></div><div>JOSM / téléchargement du cadastre vectorisé HS ?</div><div>par Maférik49 » mar. mai 19, 2020 10:06 pm</div></div>	1	430	par Maférik49 mer. mai 20, 2020 5:49 pm
<div><div></div><div>Export des favoris Google Maps vers OpenStreetMap</div><div>par Géhéf » mar. mars 17, 2020 6:03 pm</div></div>	3	1052	par garenkreiz ven. avr. 03, 2020 5:21 pm
<div><div></div><div>Comment positionner précisément un point isolé dont on connaît les coordonnées ?</div><div>par BTH » mer. janv. 29, 2020 11:36 am</div></div>	2	1208	par OsmO jeu. janv. 30, 2020 10:41 am
<div><div></div><div>Cadastre et génération de fichier .OSM est HS ?</div><div>par Xavb22440 » ven. nov. 15, 2019 8:32 pm</div></div>	1	1554	par deuzeffe sam. déc. 28, 2019 4:41 pm
<div><div></div><div>problème import adresses cadastre</div><div>par arobl » lun. oct. 21, 2019 6:28 pm</div></div>	3	1888	par deuzeffe sam. déc. 28, 2019 4:40 pm
<div><div></div><div>Revert changeset ?</div><div>par ruboqif » dim. nov. 24, 2019 10:54 am</div></div>	3	1598	par ades222 ven. nov. 29, 2019 10:00 pm
<div><div></div><div>FANTOIR connues du Cadastre mais pas d'OSM</div><div>par Xavb22440 » ven. nov. 15, 2019 7:06 pm</div></div>	4	1198	par Xavb22440 lun. nov. 18, 2019 11:18 am
<div><div></div><div>Lieu-dit</div><div>par Franncis » sam. oct. 12, 2019 12:04 am</div></div>	4	2304	par cquest mer. oct. 23, 2019 8:01 pm
<div><div></div><div>Bâtim Capture d'écran e cadastre</div><div></div></div>	3	2356	par Lukinux

Si le mode de sélection automatique ne fonctionne pas, vous pouvez utiliser le mode manuel en saisissant des sélecteurs. Nous verrons plus avant comment faire.

Mais Extractify possède également un mode alternatif se basant sur l'URL des pages de même niveau. En effet, l'observation d'un ensemble de liens de pagination par leur survol à la souris peut vous permettre de repérer un préfixe et un palier d'incrémentation :



Forum

Forum français sur openstreetmap

Rechercher...

Raccourcis

FAQ

Inscription

Connexion

Accueil du forum

Outils avancés pour contribuer

Outils avancés pour contribuer

Nouveau sujet

Rechercher...

250 sujets

1

2

3

4

5


...

10

ANNONCES	RÉPONSES	VUES	DERNIER MESSAGE
<div> <div></div> <div>Projet du mois: les défilateurs</div> <div>par gendy54 » ven. août 28, 2020 10:44 am » dans Comment contribuer</div> </div>	0	1188	par gendy54 » ven. août 28, 2020 10:44 am

SUJETS	RÉPONSES	VUES	DERNIER MESSAGE
<div> <div></div> <div>Filtres Overpass Turbo et JOSM</div> <div>par didwg » sam. août 29, 2020 9:34 pm</div> </div>	0	876	par didwg » sam. août 29, 2020 9:34 pm
<div> <div></div> <div>Bâtiment complexe</div> <div>par didwg » mar. août 04, 2020 7:24 pm</div> </div>	0	0	par didwg » mar. août 04, 2020 7:24 pm
<div> <div></div> <div>Bâtiment complexe</div> <div>par didwg » mar. août 04, 2020 7:17 pm</div> </div>	0	0	par didwg » mar. août 04, 2020 7:17 pm
<div> <div></div> <div>Lignes superposées</div> <div>par didwg » sam. août 01, 2020 8:32 pm</div> </div>	2	341	par didwg » mar. août 04, 2020 7:12 pm
<div> <div></div> <div>[JOSM] Noeuds dupliqués sur des bâtiments</div> <div>par MatthieuLyon69 » lun. août 03, 2020 2:47 pm</div> </div>	2	421	par MatthieuLyon69 » mar. août 04, 2020 3:46 pm
<div> <div></div> <div>JOSM - Couche IGN photographie 1950-1965 ?</div> <div>par Jim005 » lun. juil. 27, 2020 12:45 am</div> </div>	0	831	par Jim005 » lun. juil. 27, 2020 12:45 am
<div> <div></div> <div>Mise en œuvre via l'API d'OpenStreetMap pour permettre la géolocalisation</div> <div>par Ubaxo21 » sam. avr. 20, 2019 1:12 pm</div> </div>	1	2262	par alexosm » sam. juin 27, 2020 2:28 pm
<div> <div></div> <div>orthophoto en opendata</div> <div>par tony.emery » mar. juin 16, 2020 10:00 am</div> </div>	1	511	par cquest » mer. juin 17, 2020 7:33 am
<div> <div></div> <div>Regroupement de relations contigües</div> <div>par gsimon » dim. juin 07, 2020 11:41 am</div> </div>	7	792	par Romain » dim. juin 14, 2020 11:03 pm
<div> <div></div> <div>Comment fusionner ces trois requêtes Overpass ?</div> <div>par percherie » mer. juin 03, 2020 1:50 pm</div> </div>	2	833	par percherie » dim. juin 07, 2020 9:30 pm
<div> <div></div> <div>TMS Cadastre pour JOSM ne fonctionne plus</div> <div>par percherie » ven. mai 29, 2020 8:49 am</div> </div>	3	741	par Romain » sam. mai 30, 2020 9:30 pm
<div> <div></div> <div>Test accès BD Ortho IGN...</div> <div>par cquest » ven. mai 27, 2016 6:10 pm</div> </div>	63	46189	par wadouk » ven. mai 29, 2020 5:23 pm
<div> <div></div> <div>Comment retrouver un changeset sur une zone très limitée ?</div> <div>par percherie » sam. mai 02, 2020 3:00 pm</div> </div>	1	1087	par percherie » ven. mai 29, 2020 8:49 am
<div> <div></div> <div>Mise à jour Bano</div> <div>par Captain47 » sam. mai 19, 2018 8:43 am</div> </div>	10	5070	par cquest » ven. mai 29, 2020 8:37 am
<div> <div></div> <div>JOSM / téléchargement du cadastre vectorisé HS ?</div> <div>par Maférik49 » mar. mai 19, 2020 10:06 pm</div> </div>	1	430	par Maférik49 » mer. mai 20, 2020 5:49 pm
<div> <div></div> <div>Export des favoris Google Maps vers OpenStreetMap</div> <div>par Géhéf » mar. mars 17, 2020 6:03 pm</div> </div>	3	1052	par garenkreiz » ven. avr. 03, 2020 5:21 pm
<div> <div></div> <div>Comment positionner précisément un point isolé dont on connaît les coordonnées ?</div> <div>par BTH » mer. janv. 29, 2020 11:36 am</div> </div>	2	1208	par OsmO » jeu. janv. 30, 2020 10:41 am
<div> <div></div> <div>Cadastre et génération de fichier .OSM est HS ?</div> <div>par Xavb22440 » ven. nov. 15, 2019 8:32 pm</div> </div>	1	1554	par deuzeffe » sam. déc. 28, 2019 4:41 pm
<div> <div></div> <div>problème import adresses cadastre</div> <div>par arobl » lun. oct. 21, 2019 6:28 pm</div> </div>	3	1888	par deuzeffe » sam. déc. 28, 2019 4:40 pm
<div> <div></div> <div>Revert changeset ?</div> <div>par ruboqif » dim. nov. 24, 2019 10:54 am</div> </div>	3	1598	par ades222 » ven. nov. 29, 2019 10:00 pm
<div> <div></div> <div>FANTOIR connues du Cadastre mais pas d'OSM</div> <div>par Xavb22440 » ven. nov. 15, 2019 7:06 pm</div> </div>	4	1198	par Xavb22440 » lun. nov. 18, 2019 11:18 am
<div> <div></div> <div>Lieu-dit</div> <div>par François » sam. oct. 12, 2019 12:04 am</div> </div>	4	2304	par cquest » mer. oct. 23, 2019 8:01 pm
<div> <div></div> <div></div> <div></div> </div>	3	2356	par Lukinux »

rum.openstreetmap.fr/viewforum.php?f=5&sid=8c5efd43c9a17c2ac3cc718d7885eaa&start=25



Forum français sur openstreetmap

Rechercher...

Raccourcis FAQ

Inscription Connexion

Accueil du forum

Outils avancés pour contribuer

Outils avancés pour contribuer

Nouveau sujet

Rechercher...

250 sujets

1 2 3 4 5 ... 10

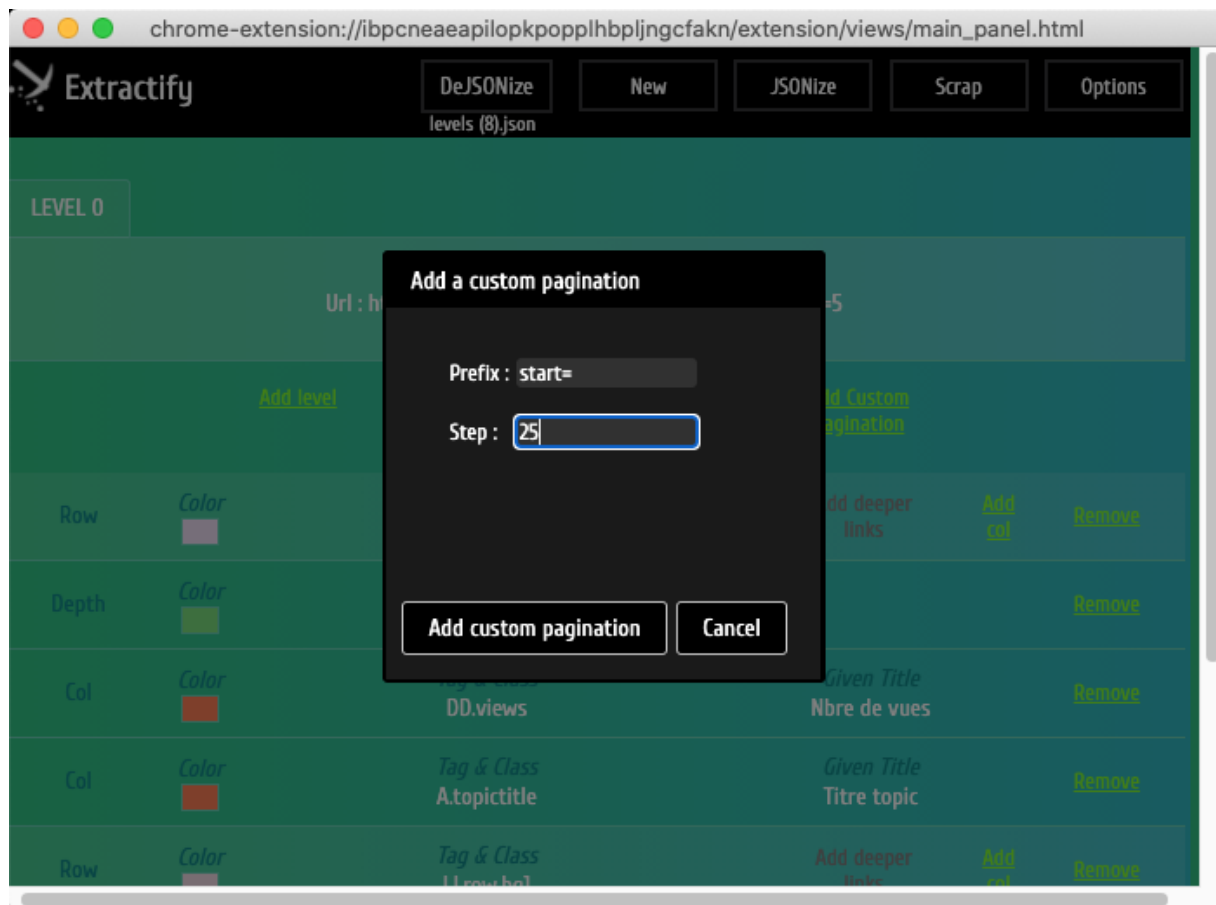
ANNONCES	RÉPONSES	VUES	DERNIER MESSAGE
<div>Projet du mois: les défibrillateurs</div> <div>par gendy54 » ven. août 28, 2020 10:44 am » dans Comment contribuer</div>	0	1188	par gendy54 » ven. août 28, 2020 10:44 am

SUJETS	RÉPONSES	VUES	DERNIER MESSAGE
<div>Filtres Overpass Turbo et JOSM</div> <div>par didwg » sam. août 29, 2020 9:34 pm</div>	0	876	par didwg » sam. août 29, 2020 9:34 pm
<div>Bâtiment complexe</div> <div>par didwg » mar. août 04, 2020 7:24 pm</div>	0	0	par didwg » mar. août 04, 2020 7:24 pm
<div>Bâtiment complexe</div> <div>par didwg » mar. août 04, 2020 7:17 pm</div>	0	0	par didwg » mar. août 04, 2020 7:17 pm
<div>Lignes superposées</div> <div>par didwg » sam. août 01, 2020 8:32 pm</div>	2	341	par didwg » sam. août 04, 2020 7:12 pm
<div>[JOSM] Noeuds dupliqués sur des bâtiments</div> <div>par MatthieuLyon69 » lun. août 03, 2020 2:47 pm</div>	2	421	par MatthieuLyon69 » mar. août 04, 2020 3:46 pm
<div>JOSM - Couche IGN photographie 1950-1965 ?</div> <div>par Jim005 » lun. juil. 27, 2020 12:45 am</div>	0	831	par Jim005 » lun. juil. 27, 2020 12:45 am
<div>Mise en œuvre via l'API d'OpenStreetMap pour permettre la géolocalisation</div> <div>par Ubaxo21 » sam. avr. 20, 2019 1:12 pm</div>	1	2262	par alexosm » sam. juin 27, 2020 2:28 pm
<div>orthophoto en opendata</div> <div>par tony.emery » mar. juin 16, 2020 10:00 am</div>	1	511	par cquest » mer. juin 17, 2020 7:33 am
<div>Regroupement de relations contigües</div> <div>par gsimon » dim. juin 07, 2020 11:41 am</div>	7	792	par Romain » dim. juin 14, 2020 11:03 pm
<div>Comment fusionner ces trois requêtes OverPass ?</div> <div>par percherie » mer. juin 03, 2020 1:50 pm</div>	2	833	par percherie » dim. juin 07, 2020 9:30 pm
<div>TMS Cadastre pour JOSM ne fonctionne plus</div> <div>par percherie » ven. mai 29, 2020 8:49 am</div>	3	741	par Romain » sam. mai 30, 2020 9:30 pm
<div>Test accès BD Ortho IGN...</div> <div>par cquest » ven. mai 27, 2020 6:10 pm</div>	63	46189	par wadouk » ven. mai 29, 2020 5:23 pm
<div>Comment retrouver un changeset sur une zone très limitée ?</div> <div>par percherie » sam. mai 02, 2020 3:00 pm</div>	1	1087	par percherie » ven. mai 29, 2020 8:49 am
<div>Mise à jour Bano</div> <div>par Captain47 » sam. mai 19, 2018 8:43 am</div>	10	5070	par cquest » ven. mai 29, 2020 8:37 am
<div>JOSM / téléchargement du cadastre vectorisé HS ?</div> <div>par Maférik49 » mar. mai 19, 2020 10:06 pm</div>	1	430	par Maférik49 » mer. mai 20, 2020 5:49 pm
<div>Export des favoris Google Maps vers OpenStreetMap</div> <div>par Géhéf » mar. mars 17, 2020 6:03 pm</div>	3	1052	par garenkreiz » ven. avr. 03, 2020 5:21 pm
<div>Comment positionner précisément un point isolé dont on connaît les coordonnées ?</div> <div>par BTH » mer. janv. 29, 2020 11:36 am</div>	2	1208	par OsmO » jeu. janv. 30, 2020 10:41 am
<div>Cadastre et génération de fichier .OSM est HS ?</div> <div>par Xavb22440 » ven. nov. 15, 2019 8:32 pm</div>	1	1554	par deuzeffe » sam. déc. 28, 2019 4:41 pm
<div>problème import adresses cadastre</div> <div>par arobi » lun. oct. 21, 2019 6:28 pm</div>	3	1888	par deuzeffe » sam. déc. 28, 2019 4:40 pm
<div>Revert changeset ?</div> <div>par ruboqif » dim. nov. 24, 2019 10:54 am</div>	3	1598	par ades222 » ven. nov. 29, 2019 10:00 pm
<div>FANTOIR connues du Cadastre mais pas d'OSM</div> <div>par Xavb22440 » ven. nov. 15, 2019 7:06 pm</div>	4	1198	par Xavb22440 » lun. nov. 18, 2019 11:18 am
<div>Lieu-dit</div> <div>par François » sam. oct. 12, 2019 12:04 am</div>	4	2304	par cquest » mer. oct. 23, 2019 8:01 pm
	3	2356	par Lukinux »

rum.openstreetmap.fr/viewforum.php?f=5&sid=8c5efd43c9a17c2ac3cc718d7885ea418start=50

Ici, nous remarquons que la navigation entre les pages s'effectue à l'aide d'une variable fixe nommée « start » à laquelle on attribue une valeur, par le signe « = », qui s'incrémente de 25 en 25.

Il suffit alors de saisir ces informations dans Extractify pour qu'il repère ces liens. Cliquez sur « Add custom pagination » et saisissez les informations voulues :

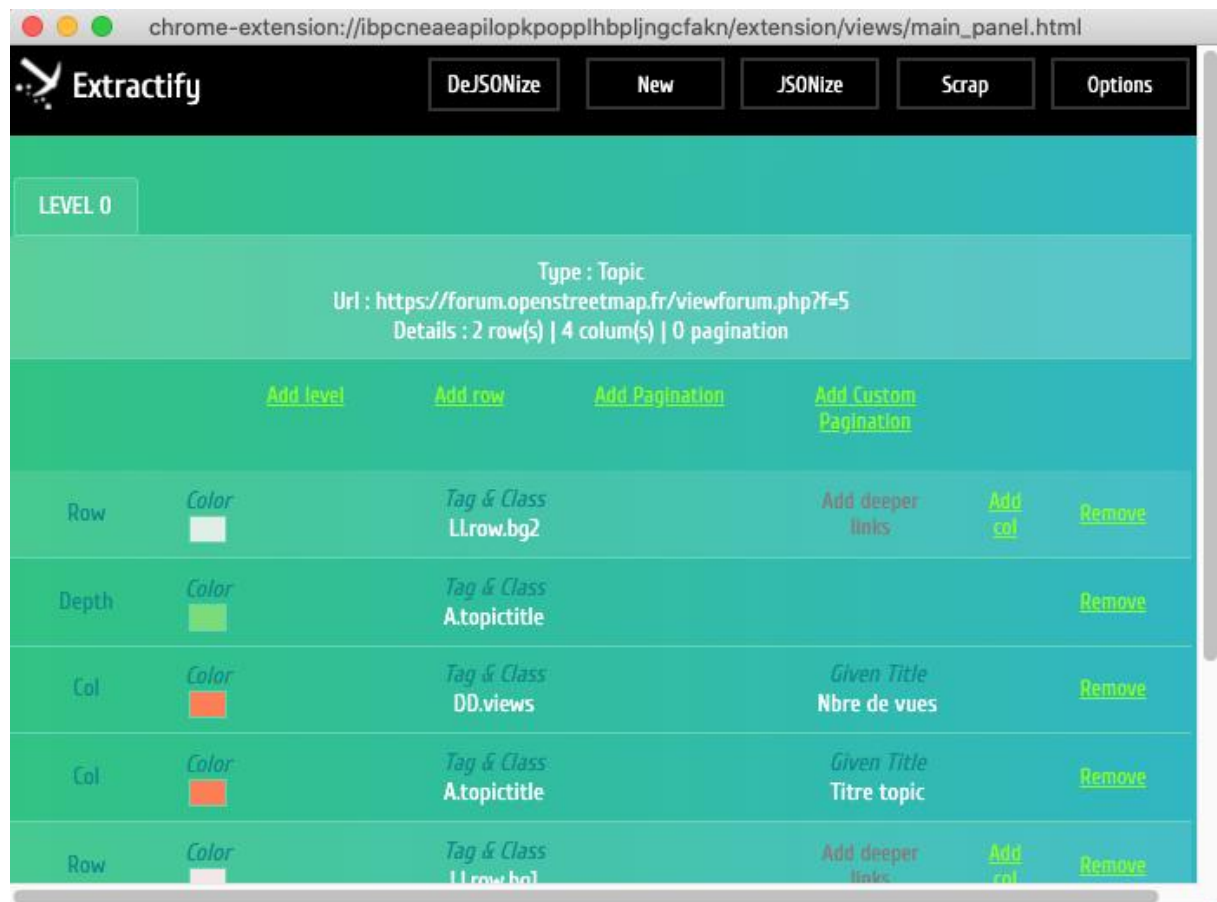


Par contre dans ce cas, les liens de pagination ne seront pas surlignés, car à aucun moment nous n'avons indiqué un élément de structure HTML permettant à Extractify de s'appuyer dessus afin d'afficher une identification visuelle.

8. Ajout d'un niveau inférieur *via* des liens de profondeur

Afin d'ajouter un niveau inférieur, vous devez tout d'abord récolter des liens vers ce niveau inférieur : ce sont des liens de profondeur.

En cliquant sur « Add deeper link », le mode opératoire de sélection est le même que pour une colonne, sauf que vous êtes restreint à la sélection d'un hyperlien. Encore une fois, vous devez répéter la manœuvre pour les 2 types de lignes :



Une fois vos liens de profondeur récoltés pour chaque ligne, un clic sur « Add level » va vous permettre d'ajouter un niveau de profondeur.

Celui-ci doit être le plus exhaustif possible, c'est à dire contenir le plus d'informations possible, dans l'esprit du proverbe « qui peut le plus, peut le moins ». Par exemple dans le cas d'un forum :

- La page de niveau inférieure est une page de messages. Nous avons vu que la structure de la présentation des topics comportait 2 types différents de lignes : il y a donc fort à parier que la présentation des messages suivra le même schéma. Il nous faudra donc choisir une page avec au moins 2 messages ;
- Toujours sur cette page de messages, nous voulons évidemment récolter le plus d'informations possibles sur les auteurs : leur date d'inscription sur le forum, leur localisation, le nombre de messages qu'ils ont envoyés, un lien vers un site web personnel...etc. : encore une fois, nous aurons besoin d'une page avec suffisamment de messages affichants chacun le maximum d'information ;

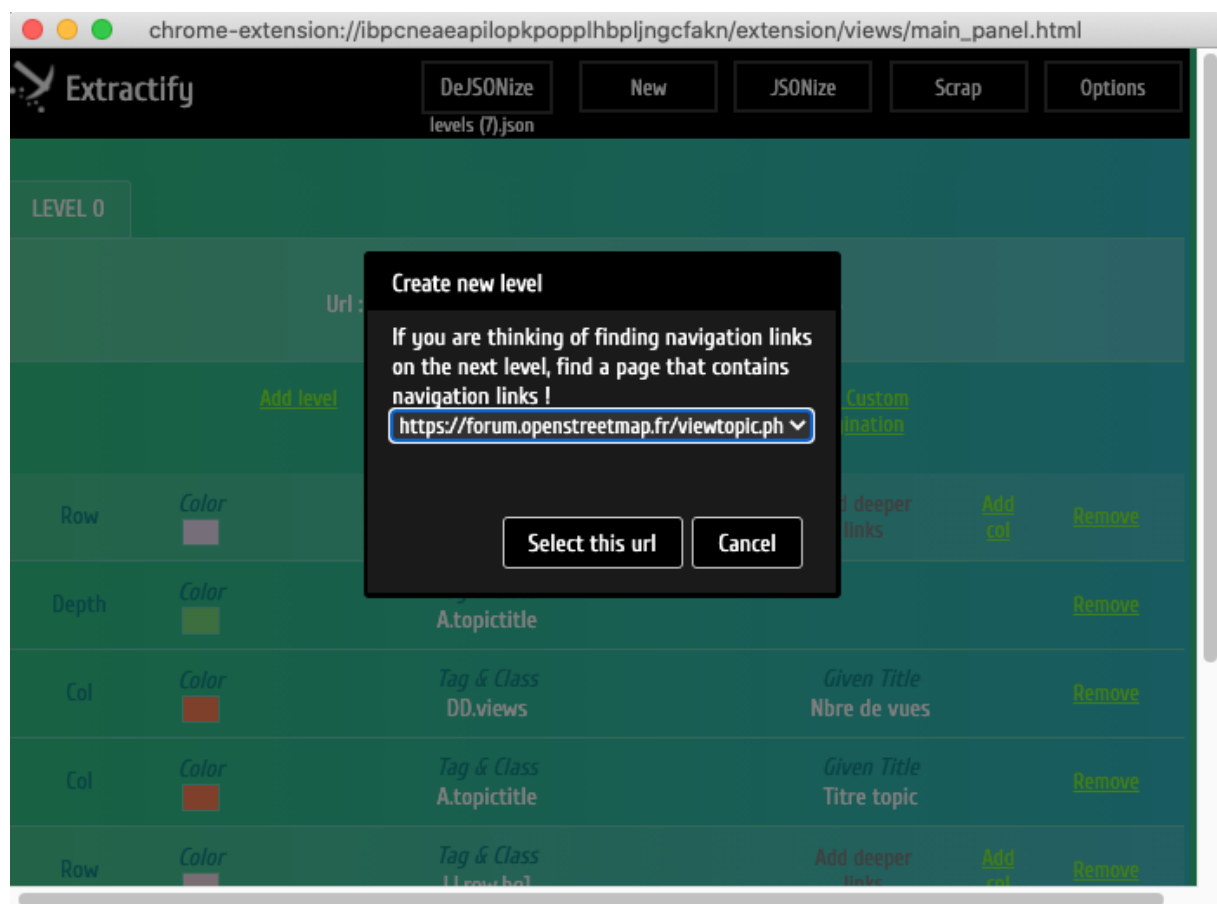
- Enfin, l'affichage des messages, comme des topics d'ailleurs, étant limité en nombre pour une meilleure lecture à l'écran, et reporté sur des liens de navigation, si nous voulons obtenir ces liens, il nous faudra là encore cibler une page avec de nombreux messages.

Une première observation visuelle de la page va nous aider à trouver la page qui pourrait remplir tous ces critères de sélection.

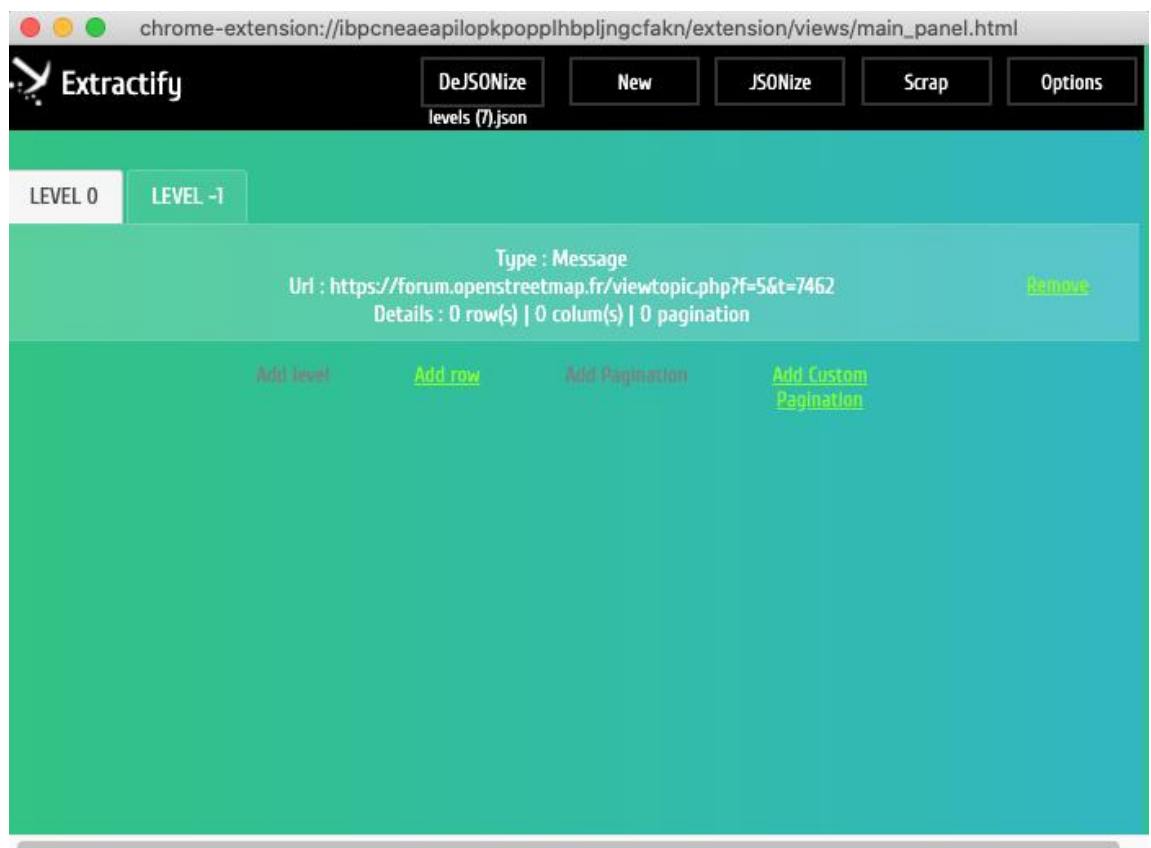
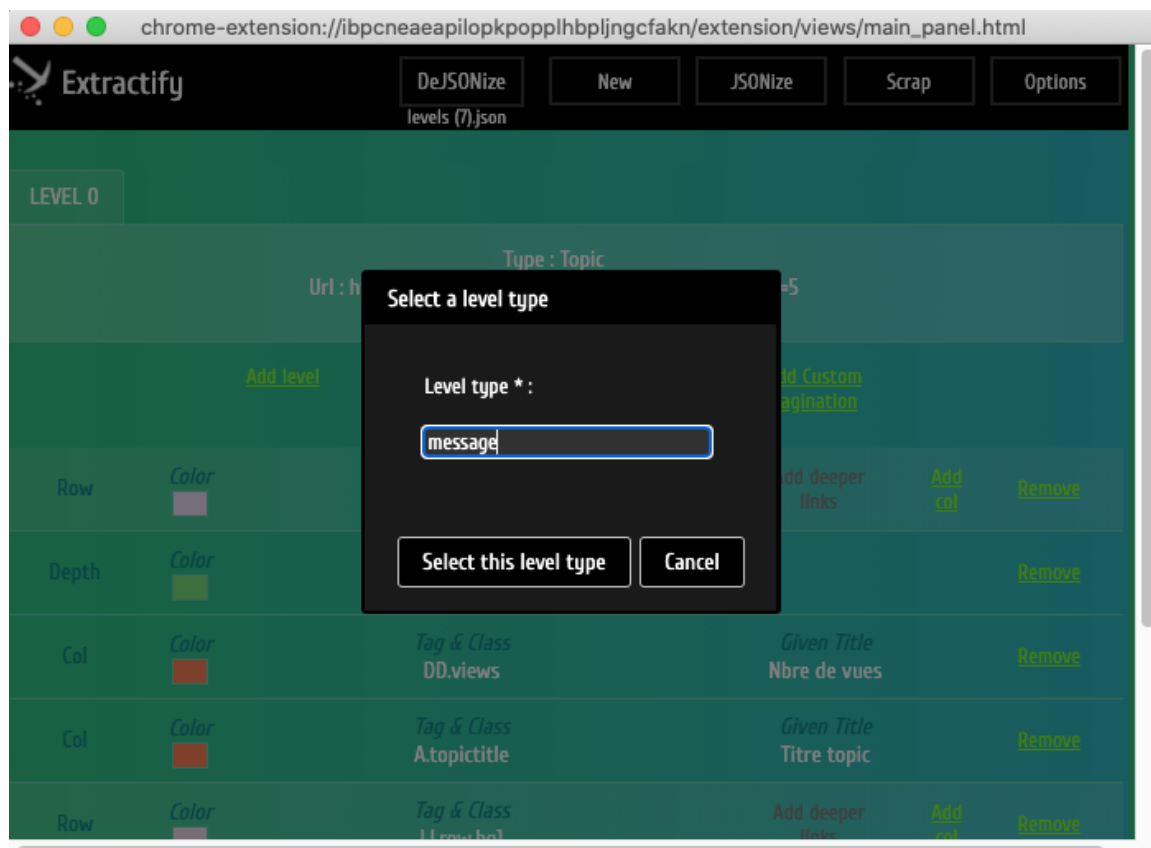
Visuellement sur la page web d'affichage des topics, nous pouvons remarquer que le topic « Test Accès BD Ortho IGN » contient un grand nombre de pages de messages. En faisant un clic droit sur la pagination des messages de ce topic puis en sélectionnant « ouvrir le lien dans un nouvel onglet », on peut voir que les 2 premières pages sont vides, mais que la 3^{ème} contient ce que nous recherchons : de nombreux messages et des liens de pagination.

Le problème c'est que cette 3^{ème} page n'est pas référencée par Extractify.

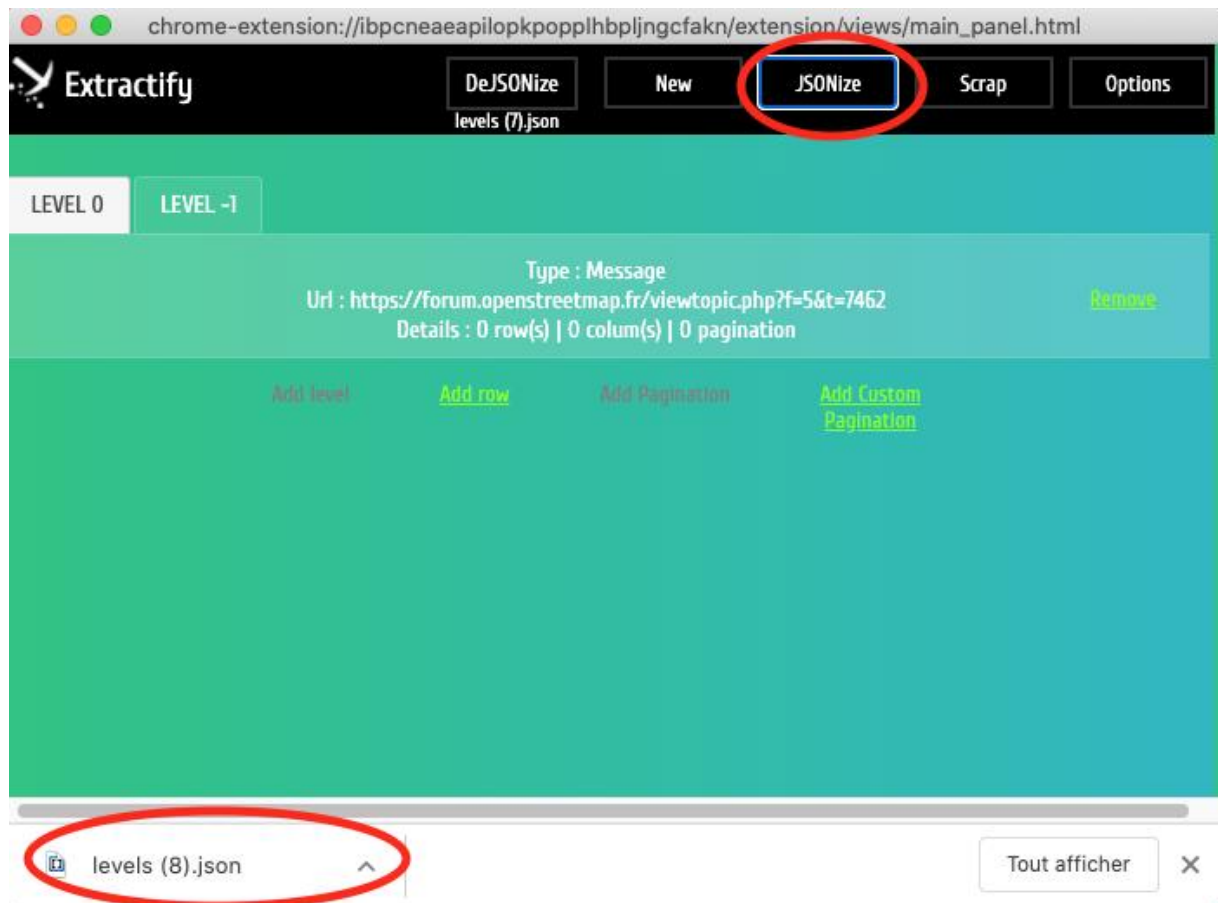
Il faut donc créer un niveau inférieur en cliquant sur « Add level » et sélectionner n'importe quel lien vers une page de niveau inférieur :



Donnez ensuite un nom à ce niveau :



Puis enregistrez votre sélection à l'aide du bouton « JSONize » :



Vous téléchargez ainsi un fichier « levels.json » qui contient la sauvegarde de votre travail de sélection.

A l'aide d'un éditeur de texte, vous pouvez alors modifier l'adresse qui va pointer vers la page de niveau inférieure. Dans notre exemple, il suffit de modifier les variables t et start de façon à cibler la page : « viewtopic.php ?t=4715&start=50 »


```

1  [
2  {
3      "id": 0,
4      "typeKey": "topic",
5      "type": "Topic",
6      "url": "https://forum.openstreetmap.fr/viewforum.php?f=5",
7      "rows": [
8          {
9              "dataType": "row",
10             "id": 0,
11             "tagClass": "LI.row.bg2",
12             "cols": [
13                 {
14                     "dataType": "col",
15                     "id": 0,
16                     "titleKey": "nbre_de_vues",
17                     "title": "Nbre de vues",
18                     "tagClass": "DD.views"
19                 },
20                 {
21                     "dataType": "col",
22                     "id": 1,
23                     "titleKey": "titre_topic",
24                     "title": "Titre topic",
25                     "tagClass": "A.topictitle"
26                 }
27             ],
28             "depth": {
29                 "dataType": "depth",
30                 "id": 0,
31                 "tagClass": "A.topictitle"
32             },
33             "color": "rgb(236, 220, 243)"
34         },
35         {
36             "dataType": "row",
37             "id": 1,
38             "tagClass": "LI.row.bg1",
39             "cols": [
40                 {
41                     "dataType": "col",
42                     "id": 2,
43                     "titleKey": "nbre_de_vues",
44                     "title": "Nbre de vues",
45                     "tagClass": "DD.views"
46                 },
47                 {
48                     "dataType": "col",
49                     "id": 3,
50                     "titleKey": "titre_topic",
51                     "title": "Titre topic",
52                     "tagClass": "A.topictitle"
53                 }
54             ],
55             "depth": {
56                 "dataType": "depth",
57                 "id": 1,
58                 "tagClass": "A.topictitle"
59             },
60             "color": "rgb(241, 219, 228)"
61         }
62     ],
63     "pagination": null,
64     "tabId": 48,
65     "someDeeperLinks": [
66         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7462",
67         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7453",
68         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7450",
69         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7138",
70         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7418",
71         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7416",
72         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7400",
73         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7412",
74         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7342",
75         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7288",
76         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7303",
77         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7240",
78         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7234",
79         "https://forum.openstreetmap.fr/viewtopic.php?f=2&t=7461",
80         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7454",
81         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7448",
82         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7445",
83         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7424",
84         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7417",
85         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=4715",
86         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=6780",
87         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7367",
88         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7304",
89         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7310",
90         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7283",
91         "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=7263"
92     ]
93 },
94 {
95     "id": 1,
96     "typeKey": "message",
97     "type": "Message",
98     "url": "https://forum.openstreetmap.fr/viewtopic.php?f=5&t=4715&start=50",
99     "rows": [ ]

```

Enregistrez ce fichier, puis importez-le dans Extractify en annulant la première boîte de dialogue d'ouverture puis en cliquant sur le bouton « DeJsonize ». Vous obtenez bien comme niveau inférieur la 3^{ème} page de messages du topic « Test Accès BD Ortho IGN » qui affiche de nombreux messages et des liens de navigations entre pages de messages.

Au sein de ce nouveau niveau, vous pouvez ainsi recommencer le processus de sélection de lignes et de colonnes.

Les lignes ne posent pas de problèmes.

Les colonnes de date de message et de contenu de message ne posent pas non plus de problèmes particuliers.

Par contre, en ce qui concerne l'auteur du message, il va falloir faire attention et utiliser l'inspecteur pour pouvoir prendre en compte tous les cas de figures :

- Le nom : les noms des auteurs ne sont pas encapsulés de la même manière selon les lignes. Une inspection le confirme : l'attribut « class » peut prendre la valeur « username » ou « username-coloured ». Il faudra le prendre en compte en ciblant les deux types de classes.
- Le rang : certains auteurs en ont un, d'autres pas
- Le nombre de messages et la date d'inscription sont communs à tous, pas de problèmes
- La localisation : là encore, certains l'affichent, d'autres pas
- Pour le contact, malheureusement Extractify n'extrait que le texte, et pas les liens. Ça sera pour une future mise à jour !

Donc pour chaque ligne, nous allons devoir cibler toutes ces colonnes, même si elles n'existent pas pour chaque message. Dans ce dernier cas, nous utiliserons le mode de sélection manuel, soit en simplement copiant-collant, soit en saisissant directement dans l'interface l'élément HTML accompagné de son attribut « class » que nous souhaitons cibler.

Le nom de l'auteur :

- On commence par sélectionner automatiquement le nom de l'auteur en bleu : pas de soucis
- Pour les noms d'auteurs en rouge, on a repéré dans l'inspecteur qu'ils étaient encapsulés dans une balise <a> accompagnée de la classe « username-coloured ».

C'est à ce moment-là que l'on va utiliser les sélecteurs. Ce sont des mots-clés qui permettent de désigner une catégorie d'éléments de la page.

https://developer.mozilla.org/fr/docs/Web/CSS/S%C3%A9lecteurs_CSS

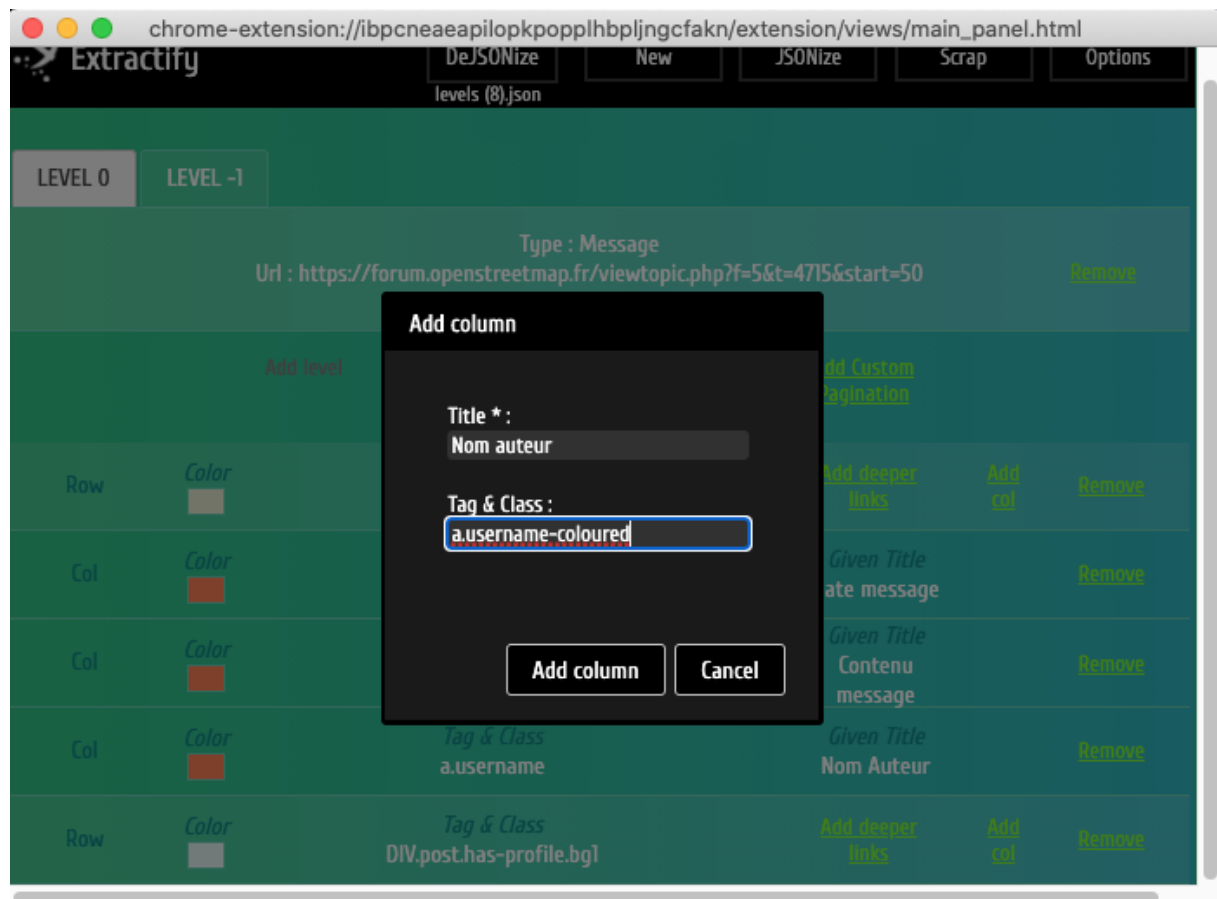
Comme vous le voyez sur cette page, il y a différents types de sélecteurs, que nous allons pouvoir utiliser pour mieux cibler ce que l'on veut extraire.

Extractify permet aussi d'utiliser des sélecteurs jquery, qui est une bibliothèque javascript, et qui étend en quelques sortes les fonctionnalités des sélecteurs CSS, en permettant par exemple la sélection de textes à partir des mots qu'ils contiennent.

D'une manière générale, lorsque le mode de reconnaissance automatique d'Extractify n'arrive pas à sélectionner ce que vous voulez, le mode manuel par sélecteur vous sauvera !

Dans le cas présent, on va donc utiliser un sélecteur de type, la balise « a », auquel on va ajouter le nom de la classe ciblée, « username-coloured ».

On ajoute donc une colonne, en saisissant le titre, puis en saisissant le nom de la balise accompagnée d'un point et du nom de la classe visée « a.username-coloured », le point étant le caractère signifiant en CSS « class » :



Sur le premier message de la page de messages, la balise <a> accompagnée de la classe « username-coloured » n'existant pas, elle ne sera pas surlignée. Mais si l'on « scroll » plus bas dans la page, on s'aperçoit que l'on a 1 message concerné, voire plusieurs !

Et on recommence de la même manière pour :

- le rang :

The screenshot shows a web application interface with a table and a modal dialog. The browser address bar indicates the URL: `chrome-extension://ibpcneaeapilopkpopplhbpljngcfakn/extension/views/main_panel.html`.

The interface has two tabs: **LEVEL 0** and **LEVEL -1**. The main content area displays a table with the following structure:

Type : Message		Url : <code>https://forum.openstreetmap.fr/viewtopic.php?f=5&t=4715&start=50</code>		Details : 2 row(s) 4 colum(s) 0 pagination		Remove	
Row	Color						
Col	Color						
Col	Color						
Col	Color						
Col	Color						
Col	Color						
Row	Color						

A modal dialog titled **Add column** is open, showing the following fields:

- Title * :** `Profil rank`
- Tag & Class :** `dd.profile-rank`

The dialog has two buttons: **Add column** and **Cancel**.

- le nombre de messages

chrome-extension://ibpcneaeapilopkpopplhbjngcfakn/extension/views/main_panel.html

LEVEL 0 LEVEL -1

Type : Message
 Url : https://forum.openstreetmap.fr/viewtopic.php?f=5&t=4715&start=50
 Details : 2 row(s) | 4 colum(s) | 0 pagination

Add level

Add column

Title * :
 Nbre messages

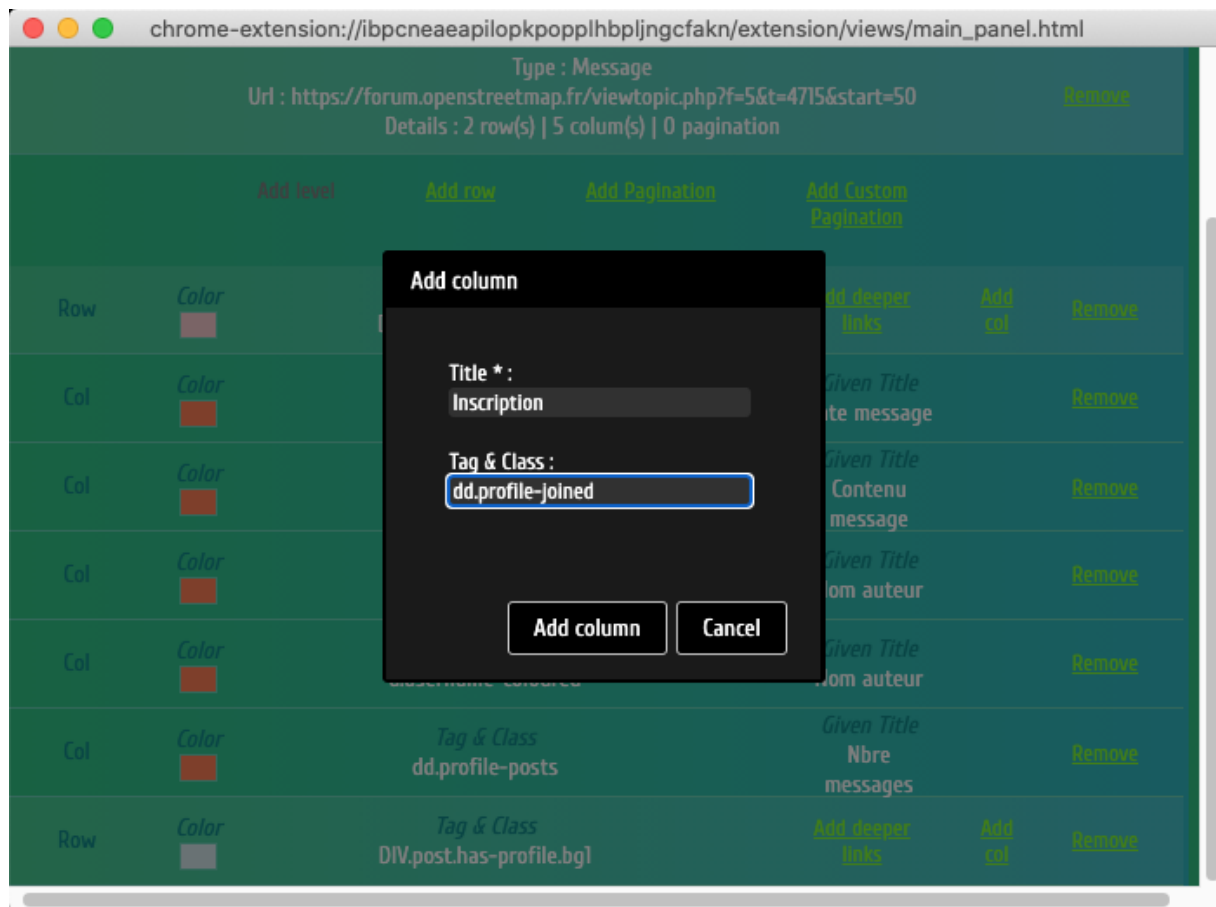
Tag & Class :
 dd.profile-posts

Add column Cancel

Row	Color	Tag & Class	Given Title	Remove
Col	Color	dd.profile-posts	ite message	Remove
Col	Color		Contenu message	Remove
Col	Color		Nom auteur	Remove
Col	Color	a.username-coloured	Nom auteur	Remove
Row	Color	DIV.post.has-profile.bg1		

Add deeper links Add col Remove

- la date d'inscription



Lorsque vous avez sélectionné toutes les colonnes pour le premier type de ligne, vous faites la même chose pour l'autre type. Vous pouvez à ce moment-là copier/coller les sélecteurs pour aller plus vite !

9. Enregistrement de la sélection

Lorsque vous avez terminé, vous pouvez enregistrer votre travail de sélection en cliquant sur « JSONize » : cela vous permet d'enregistrer les paramètres de votre sélection et de pouvoir le charger ou le modifier ultérieurement.

10. Scraping

Enfin, cliquez sur « Scrap » pour scraper les données encapsulées dans les balises que vous avez sélectionnées. A la fin du processus, le fichier est automatiquement ajouté à vos téléchargements.

11. Format de sortie

Le fichier de sortie est au format json.

Que fait-on avec ce fichier json ?

- La plupart des tableurs vont pouvoir les importer, puis vous pourrez ensuite travailler dessus, les exporter en bdd ...etc.
- D'autres logiciels type OpenRefine font aussi très bien ce travail d'importation
- Pour les données strictement conversationnelles, les données de forums de discussion, vous pouvez utiliser un autre outil que j'ai développé, qui s'appelle L@ME : <https://github.com/fredericvergnaud/lame>

12. Annexes

Extractify supporte les sélecteurs identiques à CSS (ou jquery) :

Aperçu :

- `tagname`: find elements by tag, e.g. `a`
- `ns|tag`: find elements by tag in a namespace, e.g. `fb|name` finds `<fb:name>elements`
- `#id`: find elements by ID, e.g. `#logo`
- `.class`: find elements by class name, e.g. `.masthead`
- `[attribute]`: elements with attribute, e.g. `[href]`
- `[^attr]`: elements with an attribute name prefix, e.g. `[^data-]` finds elements with HTML5 dataset attributes
- `[attr=value]`: elements with attribute value, e.g. `[width=500]` (also quotable, like `[data-name='launch sequence']`)
- `[attr^=value]`, `[attr$=value]`, `[attr*=value]`: elements with attributes that start with, end with, or contain the value, e.g. `[href*=path/]`
- `[attr~=regex]`: elements with attribute values that match the regular expression; e.g. `img[src~=(?i)\.(png|jpe?g)]`

Combinaisons :

- `el#id`: elements with ID, e.g. `div#logo`
- `el.class`: elements with class, e.g. `div.masthead`
- `el[attr]`: elements with attribute, e.g. `a[href]`
- Any combination, e.g. `a[href].highlight`
- `ancestor child`: child elements that descend from ancestor, e.g. `.body p` finds `p` elements anywhere under a block with class "body"
- `parent > child`: child elements that descend directly from parent, e.g. `div.content > p` finds `p` elements; and `body > *` finds the direct children of the body tag
- `el, el, el`: group multiple selectors, find unique elements that match any of the selectors; e.g. `div.masthead, div.logo`