

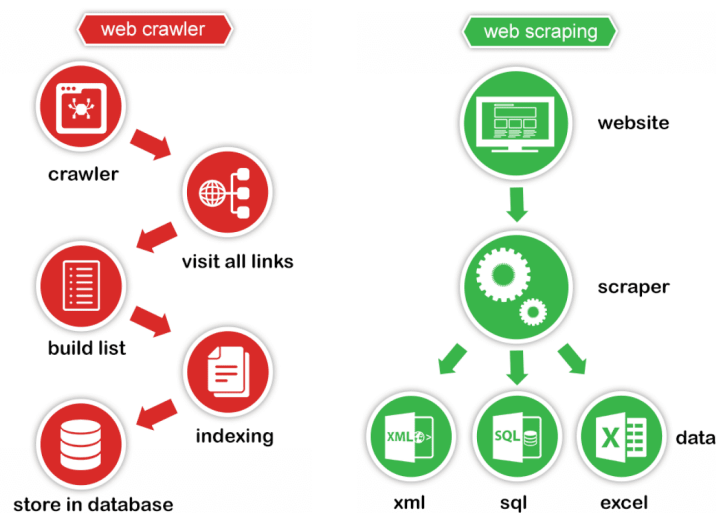
Web scraper

en c#

Qu'est-ce qu'un web scraper ?	2
Est-ce légal ?	2
Quand et pourquoi utiliser un web scraper ?	3
Fonctionnement	3
Cas concret	4
situation :	4
Préparation	4
Élaboration du scraper	7
Résultat	8
Conclusion	9

Qu'est-ce qu'un web scraper ?

Un web scrapeur peut être défini comme un robot d'exploration web. l'objectif est de récupérer les données d'un site en particulier (contrairement au crawler, il est paramétré pour un site en particulier).



Est-ce légal ?

Oui, cette pratique est dans un flou juridique, il faut donc veiller à ne pas réduire les performance du site visé (DDOS, ou ralentir le site due à un nombre de requêtes conséquent).

L'objectif est de récolter des informations, cependant il faut faire attention à l'usages des données.

Selon le site islean-consulting.fr :

“Le sujet est même éminemment complexe. Voici une mise en situation pour s'en convaincre : un internaute ou une entreprise produit et publie sur internet un ensemble d'articles. Les articles sont scrapés par un tiers et re-publiés sans modification.

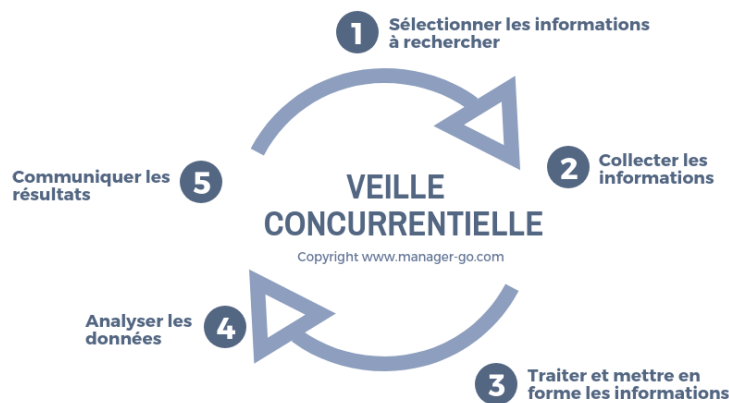
Dans ce cas précis le droit d'auteur entre en oeuvre en France et dans la grande majorité des pays du monde. Mais ce droit varie entre les pays. En France et en Europe, le droit *sui generis* stipule qu'un “investissement substantiel” doit être réalisé pour qu'une base de données [notamment de contenus] soit protégée par le droit d'auteur.

De ce fait, le scraping et la restitution d'une base de donnée scrapée sont en violation du code de propriété intellectuelle [Article L342-1] sous réserve que la transformation réalisée sur les données ne soit pas suffisamment substantielle pour justifier que la nouvelle base soit elle-même *sui generis* c'est à dire « de son propre genre ».”

Quand et pourquoi utiliser un web scraper ?

Selon le site rgdesign.fr/

“L'intérêt principal du web scraping est de pouvoir récolter du contenu sur un site web, qui ne peut être copié collé sans dénaturer la structure même du document. Ainsi cette technique est souvent utilisée dans le cadre d'une veille concurrentielle, notamment sur des sites e-commerce.”



Fonctionnement

Il y a plusieurs moyens de pratiquer le scraping :

- manuellement (une personne fait des recherches et effectue généralement des copiés collés de ce qui l'intéresse).
- automatiquement (un script qui analyse la page web et récupère automatiquement les données souhaitées).

La méthode qui nous intéresse est évidemment l'automatique.

Cas concret

situation :

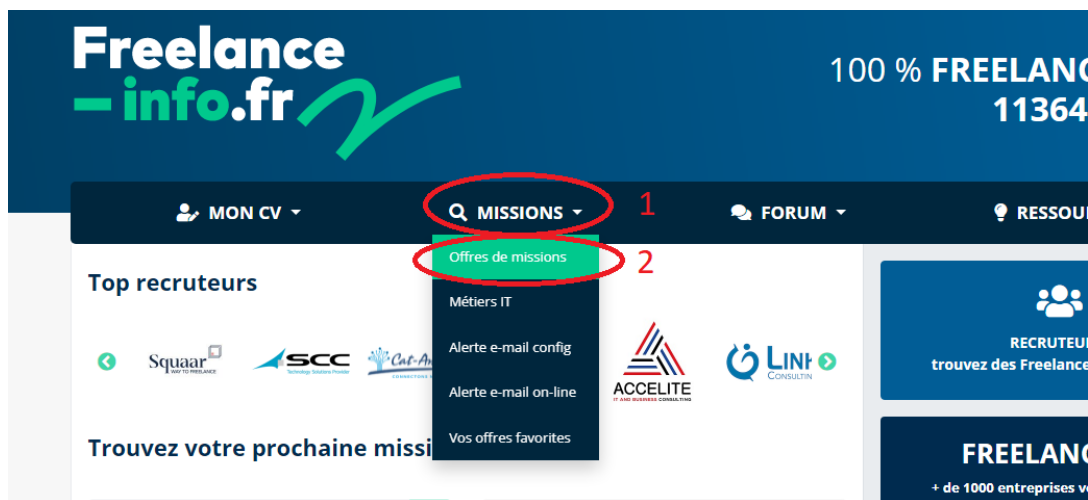
J'aimerais être en freelance, mais le temps à rechercher, lire et sélectionner les annonces de missions est trop long. J'effectue ma recherche sur le site www.freelance-info.fr et souhaite automatiser cette tâche grâce au scraping.

1) Préparation

Pour répondre à notre problématique, nous commençons par analyser notre cible. Dans un objectif de simplification, je n'aborderais pas le sujet du stockage des données récoltées et n'expliquerai pas l'entièreté du processus (l'objectif étant de comprendre le principe et l'utilité du scraping).

À l'arrivée sur le site, il y a déjà deux actions à effectuer pour accéder aux offres de missions :

- missions
- offres de missions





En observant, nous pouvons voir que nous avons accès aux missions grâce à l'URL suivante :

<https://www.freelance-info.fr/missions?page=1>

Cela réduit déjà les actions à effectuer par le scraper, nous pourrions donc charger notre page et commencer notre scraping à partir de cet URL.

La page des mission se présente ainsi, il faut donc analyser cette page grâce à l'inspecteur d'élément du navigateur (f12 sur chrome) :

Tri par pertinence Tri par date	
<div>12345>>></div>	
MANAGER DE TRANSITION FREELANCE Saint-Denis 11/03/2022 Compréhension des métiers Intégration (ingénieur de production, responsable de production)La mission consiste à définir un modèle opérationnel de management pour les 3 équipes intégration : Elaboration d'un plan de charge, d'un suivi avec les responsables de département coté Solutions ... Voir plus Non lu 24 mois 850 €	Voir l'offre 
Chef de projet MOA Risque de marché Equity Montrouge 11/03/2022 Au sein de l'équipe IT Equity Risk, vous êtes un lien essentiel entre le développement et les utilisateurs qui permet d'éclairer les différents interlocuteurs sur le fonctionnement de la chaine de production des risques de marché de la ligne métier Equity.Partenaire privilégié des métiers Risque et ... Voir plus Non lu 12 mois 450-550 €	Voir l'offre
Data Manager Confirm Paris 11/03/2022 ENTREPRISEPAC, cabinet spécialisé dans le recrutement et la délégation de compétences dans le domaine de l'informatique et du digital, recherche pour son client final, établissement financier spécialiste du crédit, un Data Manager Confirmé (h/f)MISSIONPilotage des projets data :- Établir une ... Voir plus Non lu 24 mois 500-550 €	Voir l'offre 
Data Engineer Suresnes 11/03/2022 Notre client est un leader mondial et européen dans le secteur des transactions. Ils recherchent un Data Engineer ayant entre 5 et 8 années d'expérience dans le domaine. You are in charge of defining the modern distributed data lake architecture, Pipeline, and data resiliency design standards ... Voir plus Non lu 6 mois Tarif non renseigné	Voir l'offre Huxley

div#offre.roffre 727 x 179.5

33 379 CV RÉCENTS | 118 CONNECTÉS

SE CONNECTER

MANAGER DE TRANSITION FREELANCE

Saint-Denis 11/03/2022

Compréhension des métiers Intégration (ingénieur de production, responsable de production)La mission consiste à définir un modèle opérationnel de management pour les 3 équipes intégration : Elaboration d'un plan de charge, d'un suivi avec les responsables de département coté Solutions ... [Voir plus](#)

Non lu | 24 mois | 850 €

Voir l'offre

freelance

1 à 2 mois [102]

3 à 5 mois [1186]

6 à 8 mois [2778]

> 8 mois [6725]

A définir [535]

Tarifs / jour :

< 300 € [619]

Éléments

Console

Recorder

Sources

Réseau

Performances

Mémoire

Appli

Sécurité

Lighthouse

GTM/GA Debug

```

<div>
  <br>
  <div id="offre" class="roffre"> == $0
    <div class="row"> (flex)
      <div class="col-9 pb-3 pt-3">
        <div id="titre-mission">
          <a class="rtitre filter-link" href="/mission/manager-de-transition-freelance-1674449">_</a>
        </div>
        <span class="textvert9">Saint-Denis</span>
        <span class="textgrisfonce9">11/03/2022</span>
        <br>
        <div class="text-justify">
          " Compréhension des métiers Intégration (ingénieur de production, responsable de production)La mission consiste à définir un modèle opérationnel de management pour les 3 équipes intégration : Elaboration d'un plan de charge, d'un suivi avec les responsables de département coté Solutions ... "
          <a class="text-underline" href="/mission/manager-de-transition-freelance-1674449">Voir plus</a>
        </div>
        <div class="rlig_det">
          <span>Non lu</span>
          <span> | 24 mois | 850 €</span>
        </div>
      </div>
      <div class="col-3 pt-3 pb-3">_</div>
    </div>
  </div>

```

Les informations qui nous intéressent sont :

- Le titre
- le lieux
- la date
- la description
- l'entreprise
- la durée
- l'indication "Non Lue" ou "Lu"

Maintenant que nous avons l'URL à scraper, ainsi que les données ciblés, nous pouvons commencer à programmer.

2) Élaboration du scraper

Nous allons faire le scraper en C# à l'aide de la librairie HTML Agility Pack

<https://html-agility-pack.net/>

Cette librairie est spécialement conçue à cet usage et est donc relativement simple à utiliser.

Voici les étapes que je me suis fixé pour faire mon scraper. Dans mon cas, l'élaboration des schémas et MCD était déjà réalisé.

- 1) Se renseigner sur HTML Agility Pack, Asp .net, linq et where ainsi que le moyen de stockage choisi pour les données.
- 2) Faire des premiers essais simple, en console (une sorte de maquette qui ne stock aucunes données).
- 3) Trouver un moyen de stocker les données
- 4) Créer une structure web (MVC ou API) et y intégrer le scraper

Pour m'aider dans ce projet, j'ai suivi les documentation officielles de Microsoft, HTML A Agility Pack ainsi que des vidéo trouvé sur internet.

Vidéo : <https://www.youtube.com/watch?v=wbBuB7-BaXw>

Documentation HTML A Agility Pack : <https://html-agility-pack.net/documentation>

Documentation Microsoft : <https://docs.microsoft.com/fr-fr/dotnet/csharp/>

Créer un DataTable : <https://www.delftstack.com/fr/howto/csharp/create-datatable-in-csharp/>

Développé une application web en c# (MVC) :

<https://www.youtube.com/watch?v=-tZLsJEEqeU>

3) Résultat

Les filtres appliqués sont les suivant :

- Date maximale de l'annonce : 2 jours
- une liste de mots obligatoire : Développeur, JS, JavaScript, c#
- Télétravail obligatoirement

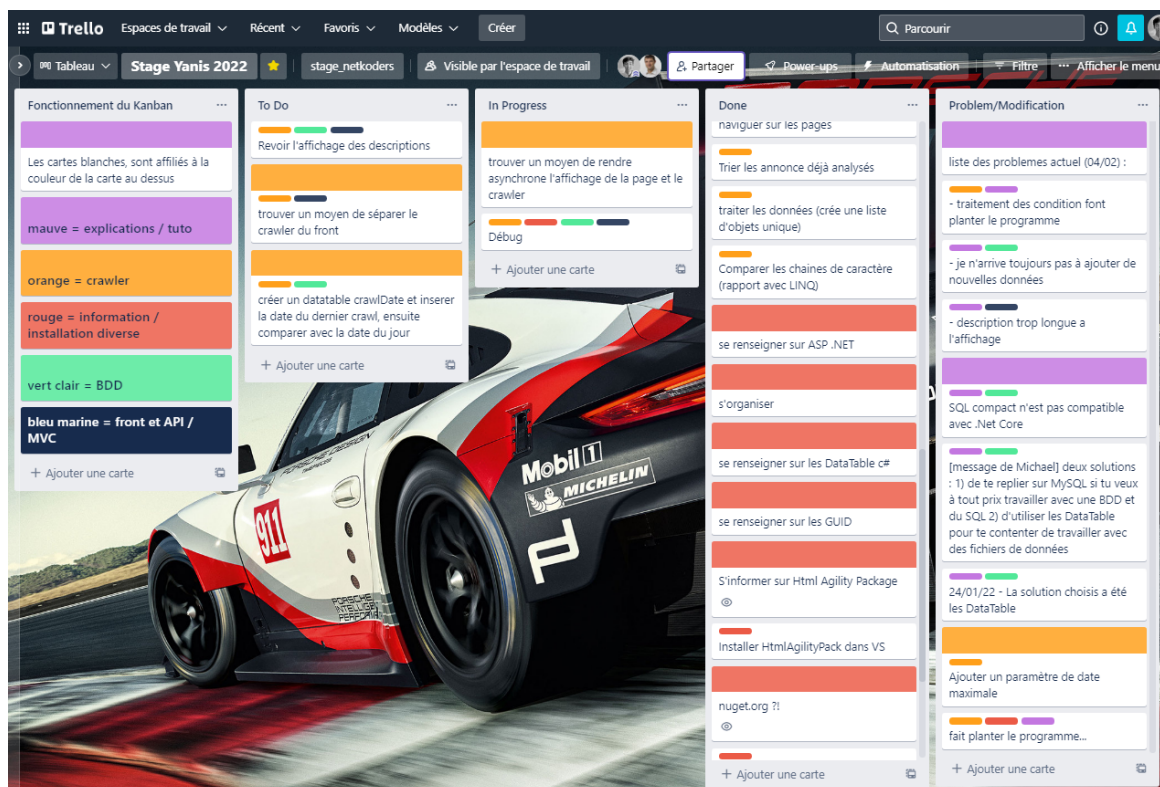
Nom d'application							
Accueil API							
<h1>ASP.NET</h1> <p>ASP.NET is a free web framework for building great Web sites and Web applications using HTML, CSS, and JavaScript.</p> Learn more »							
#	First	Last	Handle	Handle	Handle	Handle	
1	Développeur Fullstack Symfony	02/02/2022	https://www.freelance-info.fr/mission/developpeur-fullstack-symfony-1664143	Description de la mission :Développeur Fullstack, REACT (front) et Symfony (back).Assurer une veille technologiqueMéthodologie Agile (Scrum)être force de proposition dans l'amélioration et l'évolution du projetProfil recherché :Développeur Fullstack avec 3 ans d'expériencesEntreprise :Plateforme web de mise en relation	Issy-les-Moulineaux	par TRSB	6 mois 500 €
1	Développeur Fullstack Symfony	02/02/2022	https://www.freelance-info.fr/mission/developpeur-fullstack-symfony-1664143	Description de la mission :Développeur Fullstack, REACT (front) et Symfony (back).Assurer une veille technologiqueMéthodologie Agile (Scrum)être force de proposition dans l'amélioration et l'évolution du projetProfil recherché :Développeur Fullstack avec 3 ans d'expériencesEntreprise :Plateforme web de mise en relation	Issy-les-Moulineaux	par TRSB	6 mois 500 €
1	Contrôleur permanent IT h/f	02/02/2022	https://www.freelance-info.fr/mission/contrôleur-permanent-it-h-f-1664137	Description de la mission :Notre client dans le secteur de la Banque recherche un Contrôleur Permanent IT H/FDescriptif de la mission :Organisation du projetLa mission se déroulera au sein de l'équipe	Guyancourt	par LeHibou	12 mois 500-600 €

Conclusion

Un web scraper peut être utile dans certains cas (veille concurrentielle, technologique ou un marché en particulier).

En faire un est très formateur, car son développement touche beaucoup de principe C#.

J'ai eu que six semaines pour le réaliser, ce qui m'a forcé à m'organiser (chose qui pour moi était complètement abstrait au paravent).



Ce fut donc pour moi un grand pas en avant dans le monde de la programmation.