

SEO (Search Engine Optimization)

Optimisation pour les moteurs de recherche

Analyses techniques de base

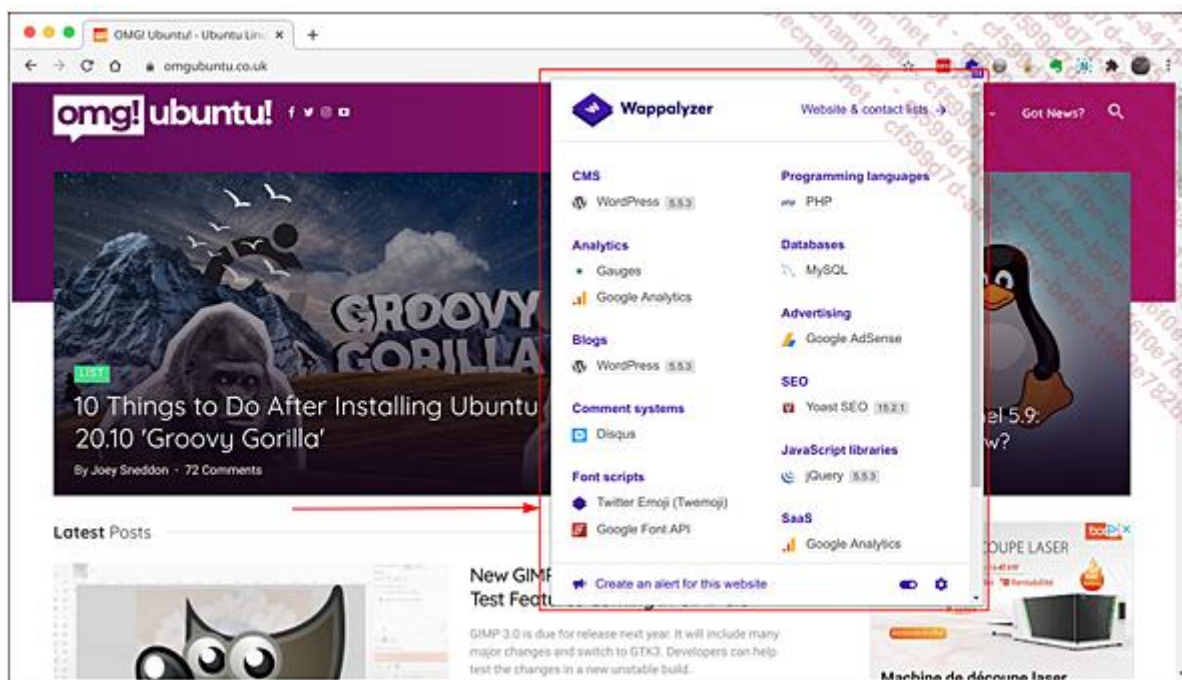
Bien sûr, la taille du site aura un impact sur vos recherches. Un site de petite taille (avec quelques centaines de pages) n'a pas les mêmes problématiques que le site d'un groupe international. Ce sera à vous d'adapter vos recherches dans ce sens.

Prenons un exemple. Les choses sont relativement plus simples pour le site d'une petite entreprise : peu d'URL, site monolingue, sous WordPress voire sous Joomla. Cela est différent d'un site dont le trafic a chuté après une migration et la fermeture de plusieurs sites satellites. Et cela est encore différent de l'analyse du site d'un groupe international, décliné en plusieurs langues (comme le chinois et le russe...) ou qui repose sur une base technique plus complexe (Drupal par exemple).

Un petit aparté : malgré la complexité de ces derniers types de sites, ce sont souvent ceux sur lesquels il est possible de mettre en place des actions correctives de grande ampleur, leurs propriétaires disposant de fonds conséquents.

Voici quelques pistes et outils qui vous donneront un aperçu des technologies employées sur un site.

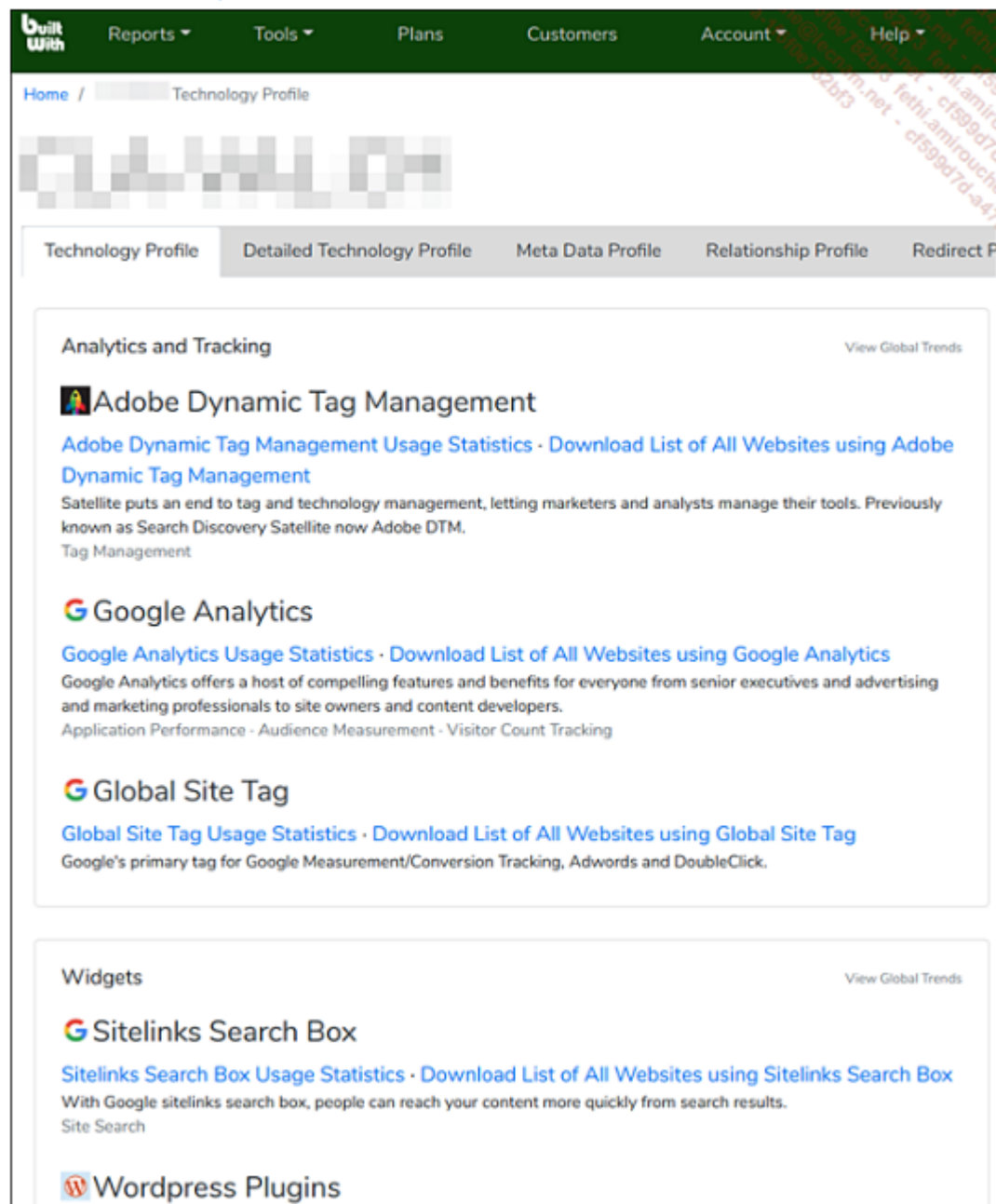
Wappalyzer : plug-in disponible sur Chrome et Firefox, il donne un aperçu de la base technique du site, notamment type de CMS, langage de programmation, type de serveur, solution analytics.



Un autre plug-in nommé **WhatRuns** est également disponible sur Chrome et Firefox (rendez-vous sur les sites des extensions des navigateurs).

- **BuiltWith.com** : ce site est une véritable mine d'or pour les apprentis chercheurs. Il présente les différentes technologies en place sur le site analysé et maintient l'historique des modifications effectuées : changements de CMS au cours des années, d'adresse IP, de code de suivi Google Analytics, etc. Il vous propose également les sites faisant une redirection depuis un nom de domaine autre. Il se peut d'ailleurs que le mandataire de l'audit ne les connaisse même pas.

Site web : <https://builtwith.com/>

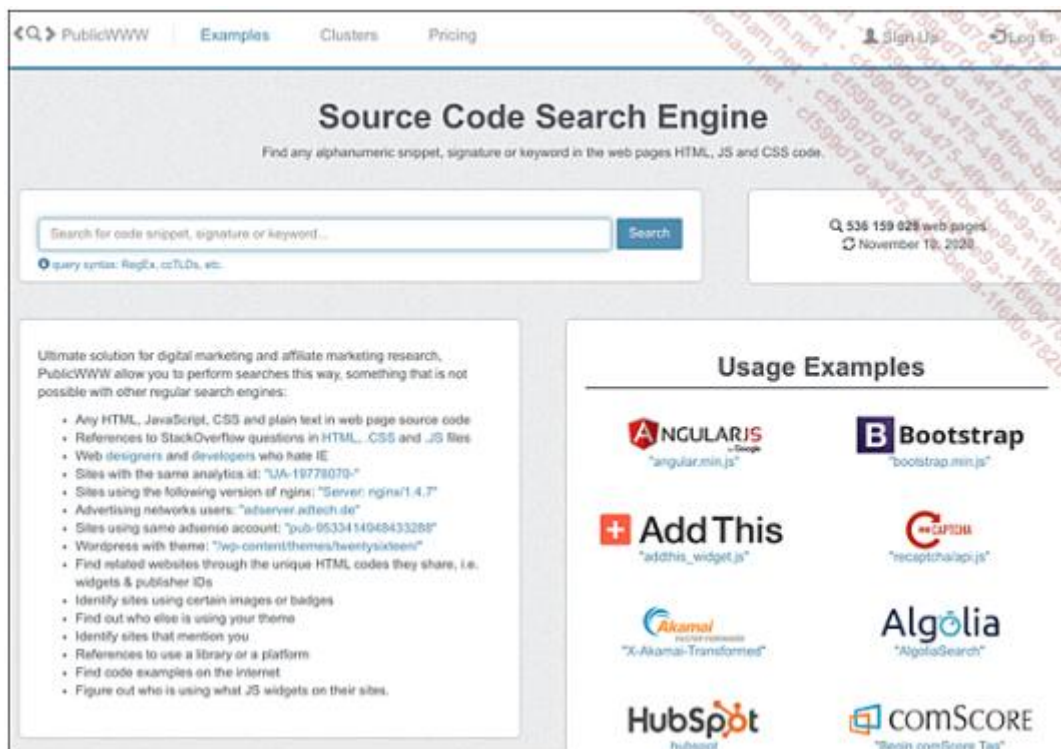


- **PublicWWW** : réservé à des recherches plus approfondies, ce site vous permet de réaliser des investigations sur des éléments non disponibles via les moteurs de recherche, mais disponibles librement dans leur code source :
 - Codes de suivi Google Analytics (pour vérifier si un code UA a pu être utilisé sur différents sites au cours de son existence).
 - Quels sites utilisent un thème WordPress particulier.

- Quels sites font mention d'une marque (ou de votre nom, si vous souhaitez analyser votre e-réputation, mais c'est un autre sujet).
- Ou encore la liste des designers ou développeurs web qui détestent Internet

Explorer : <https://publicwww.com/js/%22stupid+ie%22+filetype%3Ajs/>

Site web : <https://publicwww.com/>



- **NerdyData** : à l'image de PublicWWW, NerdyData comme son nom l'indique, vous permet également de retrouver toutes sortes d'informations sur un site ou des technologies employées. Une extension pour Chrome est disponible, mais son intérêt est limité en comparaison avec Wappalyzer ou WhatRuns par exemple. En revanche, l'intérêt de cette solution repose sur son site web qui vous permet de faire des recherches approfondies sur des codes HTML (un code HTML précis se retrouve-t-il sur plusieurs sites de votre client ?) ou sur des secteurs entiers.

Site : <https://www.nerdydata.com/>

NerdyData Technology Explorer

Search
Search for Customers

Dashboard
Save and Monitor results

Browse
Software and Industries

Monitoring
Track changes to websites

Custom Crawls
Custom Reporting

Integrations
Free Tools and Integrations

API Access
Automate and integrate data

Pricing

Support

Settings

Log Out

Websites with **all of these**:

Search for a Web Technology, or enter HTML/JS code

For example: Stripe AND Mixpanel

and

Websites with **any of these**:

Search for a Web Technology, or enter HTML/JS code

For example: Stripe OR Mixpanel

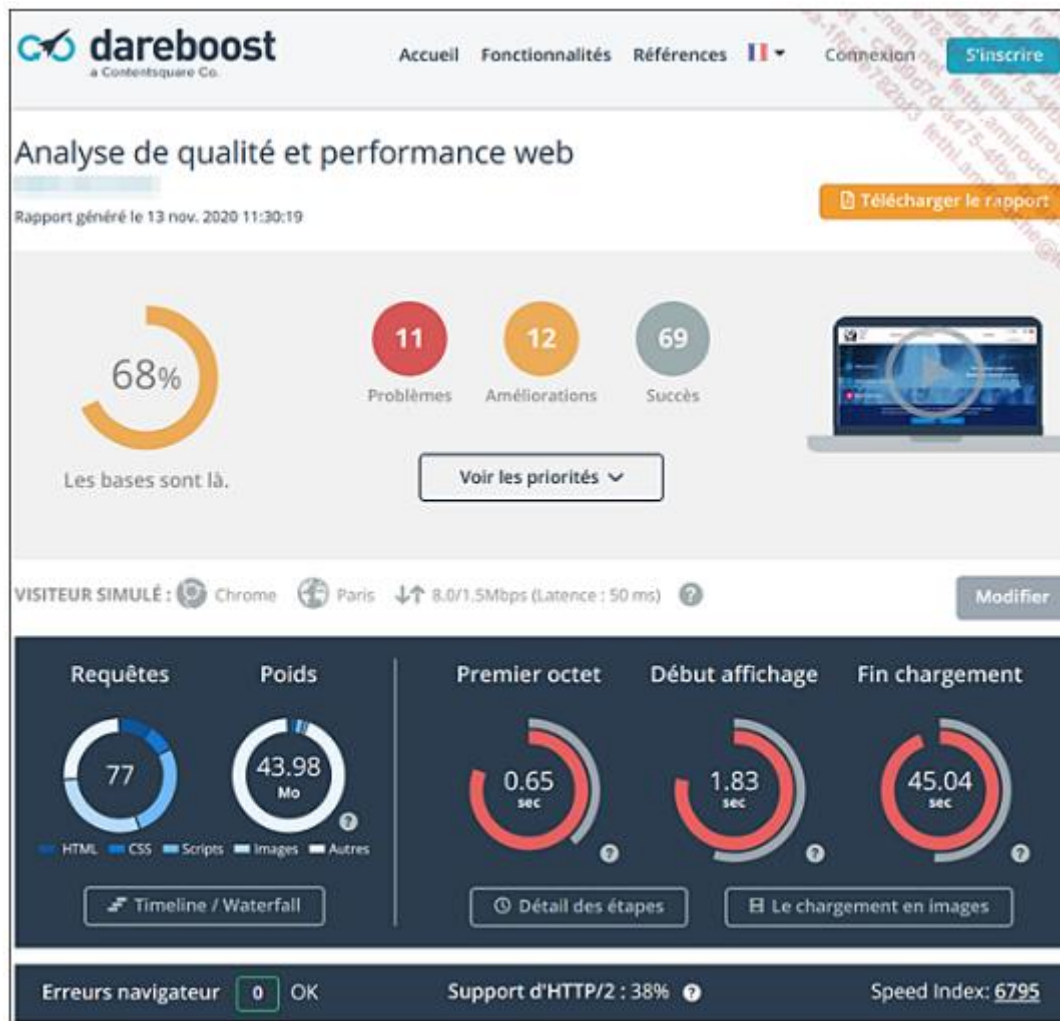
and

Websites with **none of these**:

Search for a Web Technology, or enter HTML/JS code

For example: NOT Stripe AND NOT Mixpanel

- **Dareboost** : créé initialement par une société française, Dareboost vous permet de réaliser une analyse détaillée de très nombreux points techniques et vous propose des solutions clé en main concernant des points de SEO, sécurité, UX, etc. Avec son système d'évaluation, vous pouvez vérifier quelles sont les performances d'un site. Dareboost offre d'autres services, notamment de monitoring, mais ils n'entrent pas dans le cadre de cet audit.

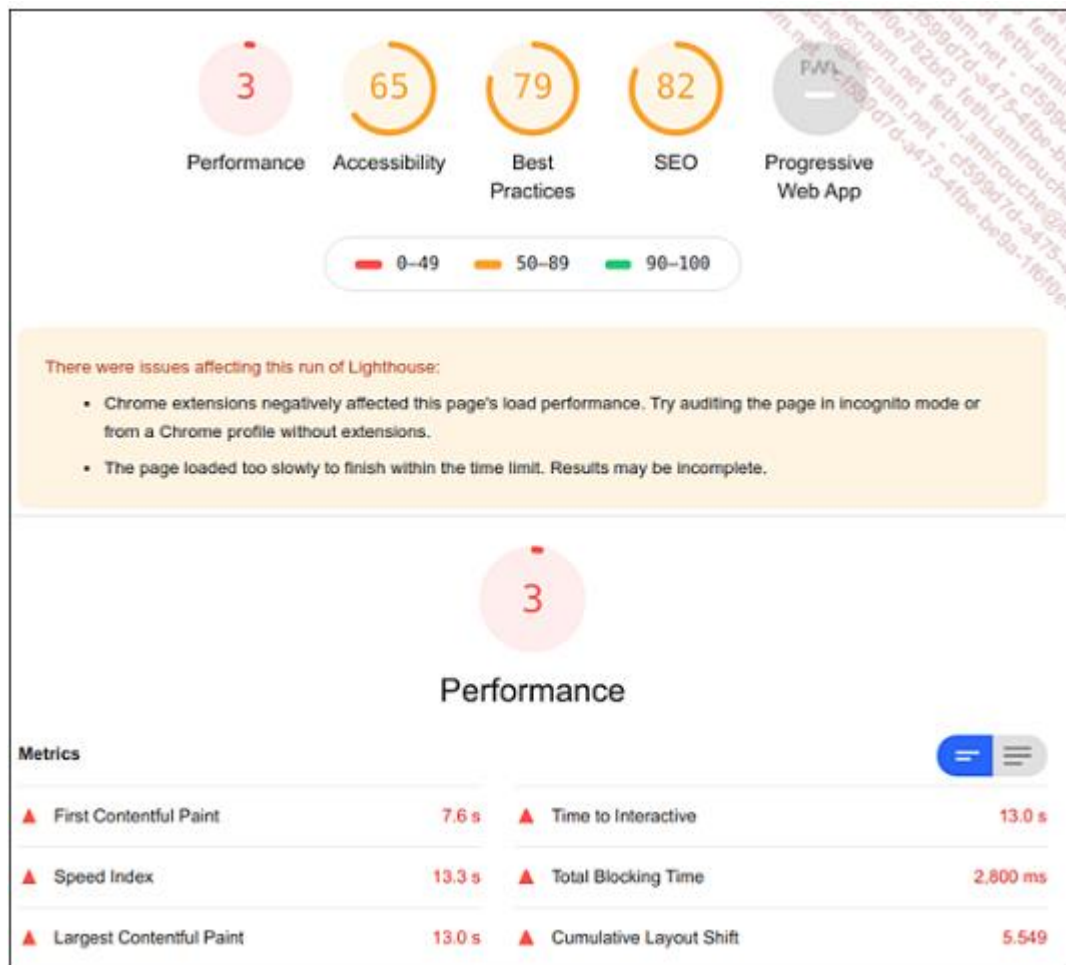


- **Google Lighthouse** : ce logiciel est très connu et aussi très visible, puisqu'il est intégré à Google Chrome (dans les outils de développeurs). Il est disponible comme plug-in pour Chrome et Firefox et accessible via le site **Google PageSpeed Insights** dit

PSI : <https://developers.google.com/speed/pagespeed/insights/>

Il peut être aussi utilisé comme module sur un serveur Node.js. Notons qu'il s'intègre à Screaming Frog. Il fournit des indications sur les performances d'un site. Elles ne sont pas à prendre au pied de la lettre, mais peuvent être intéressantes pour un aperçu « high level ». Les données peuvent être exportées au format JSON pour être exploitées plus tard grâce au

Lighthouse Report Viewer : <https://googlechrome.github.io/lighthouse/viewer/>



Pour un rendu plus concret des résultats, choisissez les options suivantes :

- **Appareil (Device)** : « **Mobile** », Google Lighthouse va ainsi simuler une connexion plus lente (la majeure partie des connexions au web se faisant via appareils mobiles).
- **Vider le cache (Clear Storage)** pour débiter les analyses avec un cache navigateur vide, comme lors d'une première connexion au site.
- **Simulate throttling** : pour simuler le changement de la page testée sur un smartphone moyen de gamme (Motorola 4G).

Ces informations sont disponibles depuis la version 80 de Google Chrome, d'autres options étaient auparavant disponibles, mais peu fiables.

Attention : le problème avec Google Lighthouse/PSI est que l'on ne connaît pas l'emplacement du serveur de test ! Et vous obtiendrez des résultats bien différents selon que vous testiez le site web à Sydney, Saint-Brieuc ou à São Paulo.

La section dédiée à l'analyse des performances de sites web vous donnera de plus amples renseignements.

- **GTmetrix et Pingdom**

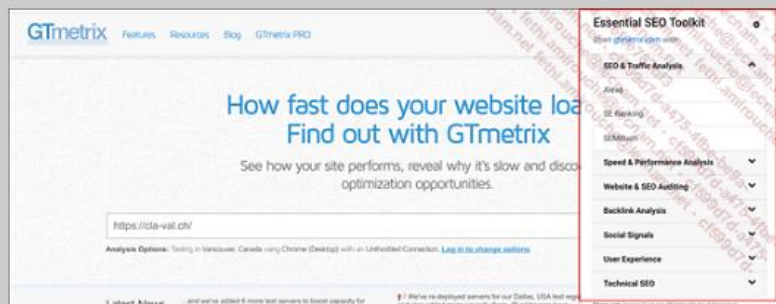
Il sera indispensable de compléter toute analyse de vitesse avec des outils tels que :

- **GTmetrix** : <https://gtmetrix.com/>
- **PingDom** : <https://tools.pingdom.com/>

Ils vous permettront de réaliser des tests depuis des serveurs situés à différents endroits du globe. Optez pour le serveur le plus proche de la clientèle que vous visez. L'utilisation de CDN (serveurs miroirs) vous permet de tester au-delà des frontières.

Utilisez l'extension Essential SEO Tool Kit pour gagner du temps dans vos analyses

Disponible pour Google Chrome, cette extension vous donne accès en deux clics aux logiciels de tests mentionnés ci-dessus ou encore à SEMrush, Alexa, Majestic, Moz et j'en passe, tout en renseignant au préalable l'URL ou le site à analyser.



Récapitulatif

Voici une liste non exhaustive des éléments que vous pouvez analyser. Je l'ai voulu aussi complète que possible, néanmoins certains points dépendent du CMS utilisé par exemple, et il est impossible de donner tous les points spécifiques à chacun.

- **Quel est le CMS employé ?** Est-il à jour ? Pour WordPress, vous pouvez vérifier le contenu de la balise meta generator.

- Y a-t-il des **sous-domaines** (potentiellement accessibles et inutiles à l'indexation) ?
- **Quel est l'hébergeur du site ?** Son adresse IP ? Sa géolocalisation (cette information est plus à titre informatif) ? Le site a-t-il changé plusieurs fois de serveur ?
- **Quelle est l'ancienneté du nom de domaine ?** Pour cela, réalisez une requête WHOIS (de nombreux sites proposent cela) ou allez jeter un œil à <https://web.archive.org/> qui « photographie » le Web depuis ses débuts.
- **Un plug-in de SEO est-il installé ?** BuiltWith vous indique notamment si Yoast SEO est présent sur les sites WordPress (le code source des pages également).
- **Le site présente-t-il des technologies bloquantes (AJAX, Flash) ?** À ce stade, vous avez déjà connaissance de ces points puisqu'ils sont nécessaires pour réaliser le crawl.
- **Le site est-il en HTTPS ?** Sachez que même si c'est le cas, des analyses ultérieures devront être entreprises pour vous assurer notamment des redirections HTTP vers HTTPS.
- **Un fichier robots.txt est-il présent** (voire un fichier robots.txt à la racine de chaque sous-domaine) ?
- **Le site présente-t-il un sitemap HTML ?**
- **Quel est son score Mobile et Desktop** sur Google PageSpeed Insights (pour la page d'accueil du site ou pour toute page importante) ?
- **Quel est son score sur Dareboost.com** (pour la page d'accueil du site ou pour toute page importante) ?
- **Le site est-il adapté aux appareils mobiles ?**
Voir **Google TestMySite** : <https://www.thinkwithgoogle.com/intl/fr-fr/feature/testmysite/>
et <https://search.google.com/test/mobile-friendly>
- **En cas d'échec du plan de redirection ou de migration**, quels sont les sites initiaux et informations particulières ? On s'intéressera aux informations susmentionnées pour chacun d'eux.

Autres outils

- **Sitespeed.io** : cette solution gratuite et open source permet d'analyser un site à la manière de Google PageSpeed Insights. Il s'agit d'un serveur qui va « passer au peigne fin » différents éléments et vous fournir un résumé détaillé. C'est une solution à réserver aux personnes à l'aise avec la ligne de commande, même si les manipulations sont plutôt simples. Vous pouvez notamment faire tourner ce service sur votre propre machine à l'aide d'un container Docker, le site vous fournit les lignes de commandes. Docker fonctionne en natif sous Linux notamment.

Voici un exemple de ligne de commande :

```
docker run --rm -v "$(pwd) : /sitespeed.io"  
sitespeedio/sitespeed.io : 15.9.0 https://URL-de-votre-page
```

Ici, il vous suffit de copier/coller cette ligne de code dans un terminal, et le logiciel va analyser l'URL demandée. Pour les plus férus d'entre vous, les données peuvent même être exportées vers des outils de data visualisation (Graphite et Graphana), qui s'adressent plutôt à des publics IT et des développeurs.

<https://www.webpagetest.org/> : encore un outil à la manière de GTMetrix ou Pingdom.com, vous offrant un grand choix d'appareils et de serveurs pour analyser la vitesse de votre site.

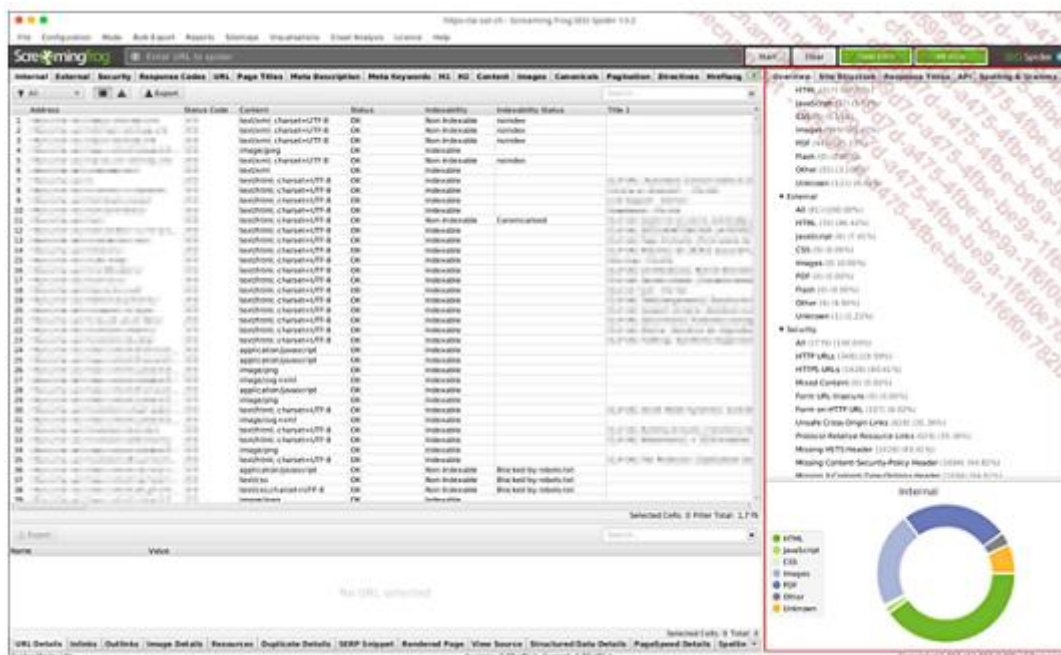
- **Mozbar (extension Chrome)** : c'est une petite extension, il en existe un très grand nombre sur le marché. Elle est notamment utile pour partir à la recherche en un clic du code source (DOM) d'une page qui se trouve dans le cache de Google. Cette extension est pratique pour vérifier si le contenu à l'air compréhensible, s'il existe ailleurs ou si la page rencontre des problèmes (vide, présence de très nombreux liens sortants, etc.).

Extraction des données de Screaming Frog

Nous allons ici nous intéresser aux résultats du crawl, aussi bien dans l'interface de Screaming Frog que dans Excel vers où nous allons exporter les données.

1. Brève analyse des données dans Screaming Frog

Pour obtenir un premier bilan de la situation du site, faites un rapide repérage des informations remontées par Screaming Frog. Les informations se trouvent dans le panneau de droite dans l'onglet **Overview**.



Vous trouverez les éléments suivants :

- **Un résumé du crawl** : nombre d'URLs, nombre d'URLs bloquées par le robots.txt, etc.
- La nature et le nombre d'**éléments internes** crawlés (HTML, CSS, JS, PDF, Flash, autres).

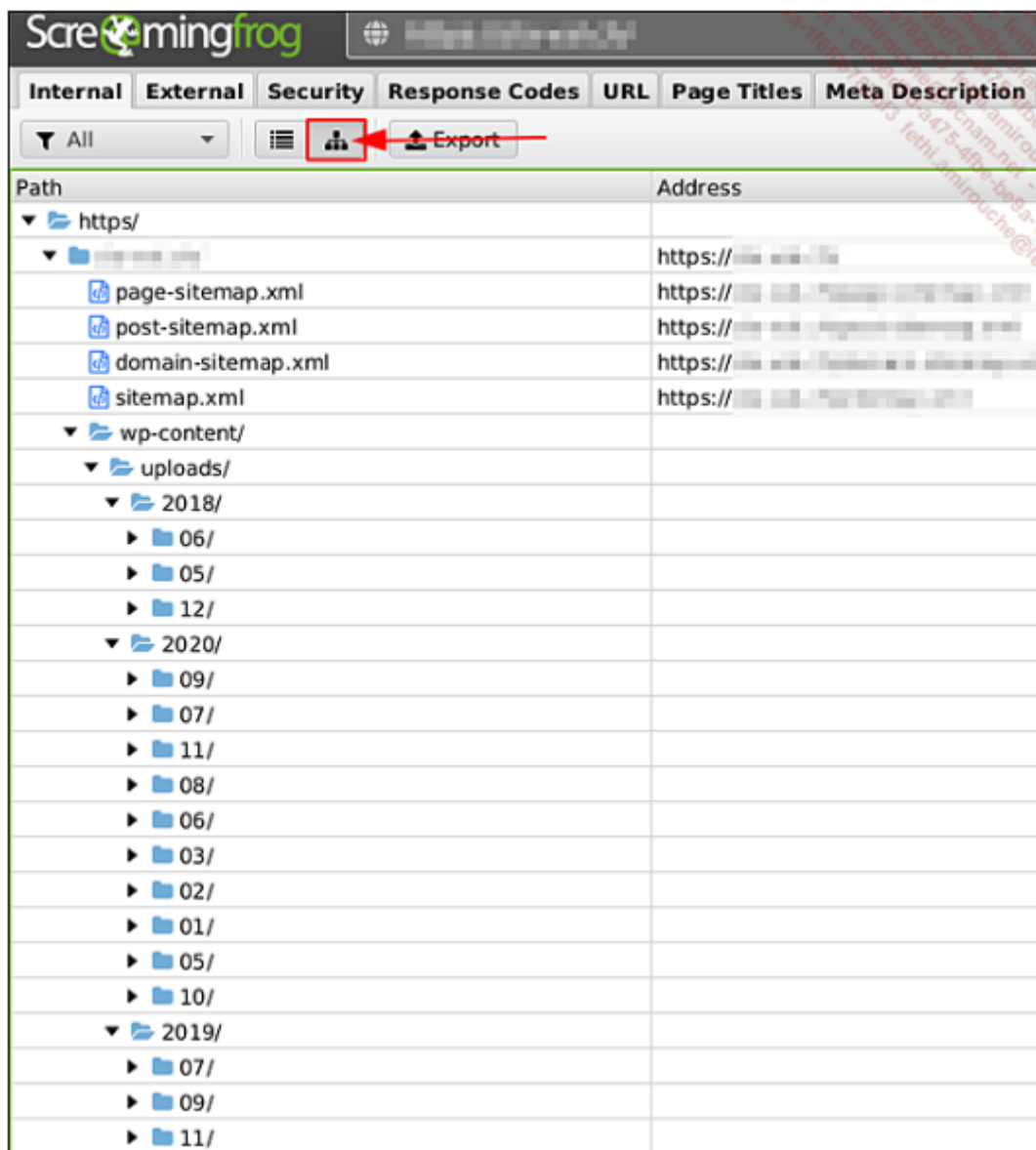
- La nature et le nombre d'**éléments externes** crawlés (HTML, CSS, JS, PDF, Flash, autres).
- Les informations relatives à la **sécurité du site** (HTTPS, HSTS, etc.).
- Les codes de réponses des pages (200, 301, bloquées par robots.txt, etc.).
- Les informations relatives aux **URLs** (présence d'underscore, de paramètres, etc.).
- Les informations sur les éléments **on-page** textuels (title, meta description, H1, etc.).
- Les informations sur les **images**.
- Les informations sur les **balises *canonical***.
- Les informations relatives à la **pagination**.
- Les détails sur les directives rencontrées sur les pages (nosnippet, nofollow, refresh, etc.).
- Les informations sur les **balises hreflang**.
- Les détails des recherches **custom** que vous auriez pu faire.

Les résultats de ces différents points dépendent du paramétrage de Screaming Frog que vous avez réalisé. En utilisant les préconisations prescrites dans la section précédente, vous obtiendrez des informations pour toutes les catégories.

Un résumé du crawl peut être obtenu en cliquant sur le menu **Reports - Crawl overview**.

La deuxième analyse que vous pouvez faire consiste à **afficher l'arborescence du site**. Vous pourrez y déceler des renseignements divers (pages HTTP bien renvoyées en HTTPS, URLs mal formées, etc.).

Pour cela, en haut à gauche de l'interface de Screaming Frog, choisissez l'onglet **Internal** et cliquez sur le bouton **Select tree table view**. Vous pouvez notamment filtrer uniquement les pages web (HTML).



2. Export des URLs

Les informations récoltées lors du crawl vont permettre de réaliser l'analyse d'une grande partie des rubriques à venir. Vous avez pour l'instant réalisé le crawl du site et sauvegardé les données du crawl en lieu sûr (fonction Export).

Nous allons ici **exporter les données** qui nous intéressent le plus : les pages web internes. Pour cela, il vous faudra choisir **HTML** dans le menu déroulant en haut à gauche puis cliquer sur le bouton **Export**.

Le logiciel vous propose d'exporter les données aux formats CSV, Excel ancienne version (format .XLS) et Excel (.XLSX). À titre personnel et par défaut, je choisis CSV afin de pouvoir utiliser les données dans d'autres logiciels si besoin, puis je convertis le fichier au format Excel (.XLSX). Pour gagner du temps, vous pouvez directement choisir ce dernier format.

Je vous propose également une nomenclature de noms de fichiers. S'agissant des pages du site de notre client, nous utiliserons le montage suivant : nomdedomaine.tld-PAGES.xlsx (ou .csv). Vous pouvez renommer la feuille de calcul actuelle contenant les URLs du site en **URL**. Enfin, nommez le tableau lui-même **URL** (cette option se trouve dans les paramètres du tableau dans l'onglet **Conception de tableau**).

Si vous devez analyser les sites des concurrents, vous procéderez de même.

À retenir : nomenclature

Nom du fichier contenant les URLs : nomdusite.tld-PAGES.xlsx.

Nom de la feuille de calcul d'URL : URL.

Faites de même pour les sites des concurrents si besoin.

3. Mise en forme des données avec Excel

Une fois les données importées dans Excel, vous pouvez remettre en forme ce tableau. Commencez en supprimant les deux premières lignes, elles ne servent à rien. Ensuite, utilisez l'outil **Mettre sous forme de tableau** et choisissez la mise en forme souhaitée.

Ces informations se retrouvent dans la partie **Overview** de Screaming Frog. Il est intéressant de compléter le tour d'horizon fait précédemment dans cette partie du logiciel par une analyse des données brutes. Votre appréciation s'affinera avec l'expérience.

Une fois ce tour d'horizon fait, mettons de côté ce document. Nous avons un aperçu de ce qui nous attend et des problèmes auxquels nous allons être confrontés. Ce document nous servira de base pour certaines des analyses à venir. Mais il est temps d'avancer dans les analyses et dans la rédaction de notre rapport d'audit pour les personnes concernées.

Analyse du crawl et de l'indexation

Ici, nous allons traiter tous les éléments qui peuvent être des freins au passage des robots (Googlebot, BingBot, etc.) et à l'indexation du site. Google Search Console nous avait déjà fourni des indications précieuses. Nous allons pouvoir leur trouver des explications et des solutions. Vous pourrez constituer votre rapport d'audit à partir des points abordés ci-après.

1. Technologies du site

Le site utilise-t-il des technologies qui bloquent le crawl ? Nous faisons référence ici aux sites conçus en AJAX/JavaScript ou ceux qui utilisent encore du Flash.

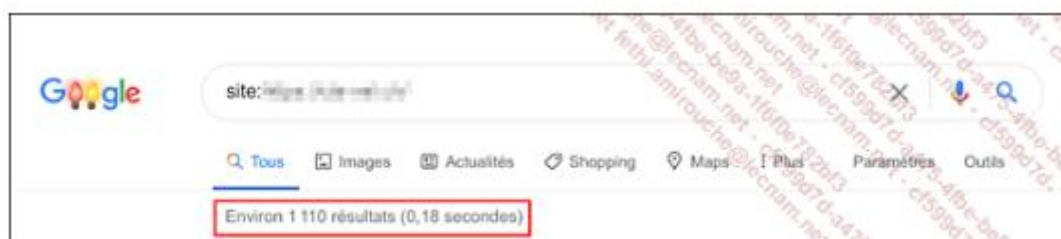
Comme expliqué précédemment, les sites en AJAX nécessitent des ressources supplémentaires pour être compris par les moteurs de recherche. Et si ces derniers ont fait des progrès en la matière, les analyses montrent que les moteurs de recherche vont moins en profondeur (en nombre de clics depuis une page de référence) que pour les sites traditionnels.

2. Faire le bilan du crawl

Dans le reporting, vous allez transmettre les données générales du crawl puis entrer dans le détail des pages qui ressortent bien et celles qui sont bloquées.

Données Google

Faites un récapitulatif des données d'indexation fournies par Google sur la requête `site:url-du-site.tld`



Données générales

Concernant les données générales, vous pouvez récapituler :

- la date du crawl,
- le nom du logiciel utilisé (Screaming Frog),
- le nombre d'URLs crawlées.

Vous pouvez notamment préciser les chiffres fournis dans le rapport Screaming Frog intitulé Crawl Overview disponible le menu **Reports** du logiciel.

Vous pouvez également fournir des renseignements sur la configuration de Screaming Frog en tenant compte du robots.txt, ou des préconisations de Google expliquées précédemment. Si vous réalisez une présentation, ces informations peuvent très bien être communiquées à l'oral.

Liste des ressources découvertes

Nous allons ici faire remonter **les différentes catégories** de ressources rencontrées par le crawler. Cet aspect va nous permettre d'apprécier le site en nous donnant un aperçu global de ses fichiers. Mais avant d'entrer dans la partie technique, un peu de théorie s'impose.

Screaming Frog utilise la classification par type **MIME** (*Multipurpose Internet Mail Extension*), qui à l'origine répertoriait les types de médias pouvant être envoyés par mail. Cette classification s'est étendue à tous les types de fichiers circulant sur Internet. C'est notamment un élément essentiel des en-têtes HTTP envoyés par les serveurs web aux navigateurs.

Cette classification se présente plus exactement sous la forme suivante : Type MIME/Encodage où :

- le type MIME est la ressource ou le type de données (ex. : text, image, application),
- l'encodage est la forme que prend ces données (ex. : *plain* pour du texte, *jpeg* pour des images, *pdf* pour des applications).

Il existe dix types MIME : application, audio, example, font, image, message, model, multipart, text et video.

Screaming Frog fournit ainsi un résumé détaillé de chaque ressource rencontrée par type MIME et encodage, qu'elle soit interne (présente sur le serveur) ou externe (ex. : API Google Maps, bibliothèques JavaScript chargées depuis des CDN).

Aperçu des ressources

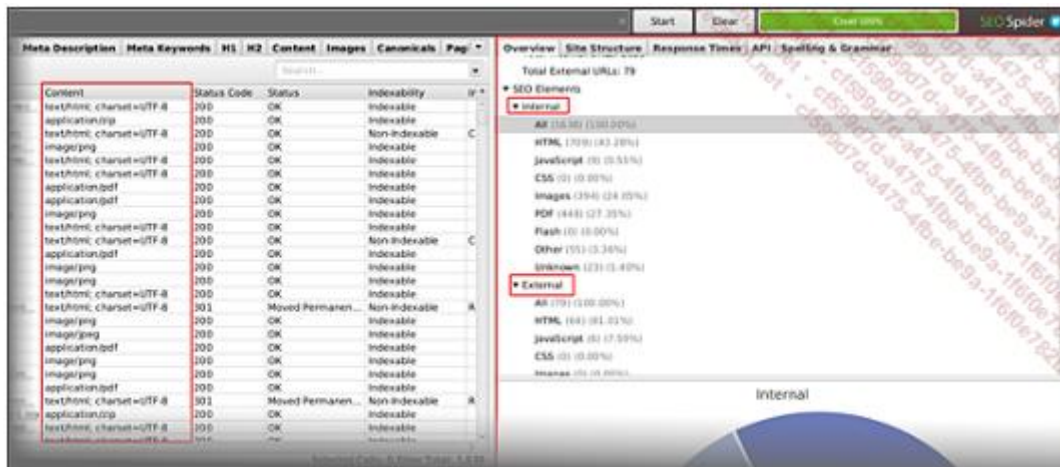
Il vous suffit de visualiser les données fournies dans l'onglet **Overview** de la colonne de droite de Screaming Frog aux rubriques **SEO Elements - Internal** et **SEO Elements - External**.

Vous retrouvez naturellement ces éléments au format texte exploitable dans le rapport **Reports - Crawl Overview**, prêts à être copiés-collés dans votre rapport d'audit.

À travers l'examen des ressources, il est possible de mieux comprendre les fonctionnements et de déceler de possibles problèmes :

- **Absence de CSS** : sont-elles internes ou chargées via du JavaScript ?
- **Présence de PDF** : sont-ils bien optimisés pour le SEO et surtout pourrait-on les remplacer par des pages web ?

- Présence de fichiers Flash
- Tous les éléments renvoient-ils un code HTTP 200 ?



3. Analyse du fichier robots.txt

Il s'agit évidemment de la première étape véritablement technique de notre analyse. Le fichier **robots.txt** est audité par les moteurs de recherche avant toute entrée en matière au niveau du crawl. Ce fichier indique quelles parties du site peuvent être accessibles à des robots.

Son intérêt est double :

- Éviter que des pages sans intérêt pour les internautes n'apparaissent dans les résultats des moteurs de recherche (SERP). C'est notamment le cas des pages d'administration des sites.
- Éviter de consommer du budget crawl de manière inutile.

Définition : Budget crawl

Les moteurs de recherche utilisent des ressources techniques pour analyser les sites web. Concernant le crawl, ils allouent un budget crawl (*Crawl Budget* en anglais) qui définit la quantité d'URL à analyser selon différents critères (taille du site, profondeur des pages, technologies du site ou fraîcheur du contenu notamment).

Des freins au crawl (par exemple le recours à de l'AJAX ou encore un site trop profond) peuvent consommer du budget crawl. Google pourrait ainsi faire le choix d'analyser uniquement les pages situées à un niveau de profondeur moins important. Ainsi, des pages ou autres ressources qui apparaîtraient comme

crawlables, mais sans intérêt pour l'utilisateur gaspilleraient du budget crawl inutilement, au détriment de pages plus importantes.

Cette notion s'applique principalement aux grands sites, mais une optimisation des sites plus modestes reste intéressante.

Ici nous devons nous assurer que le fichier robots.txt est présent et bien configuré.

Checklist d'analyse du fichier robots.txt

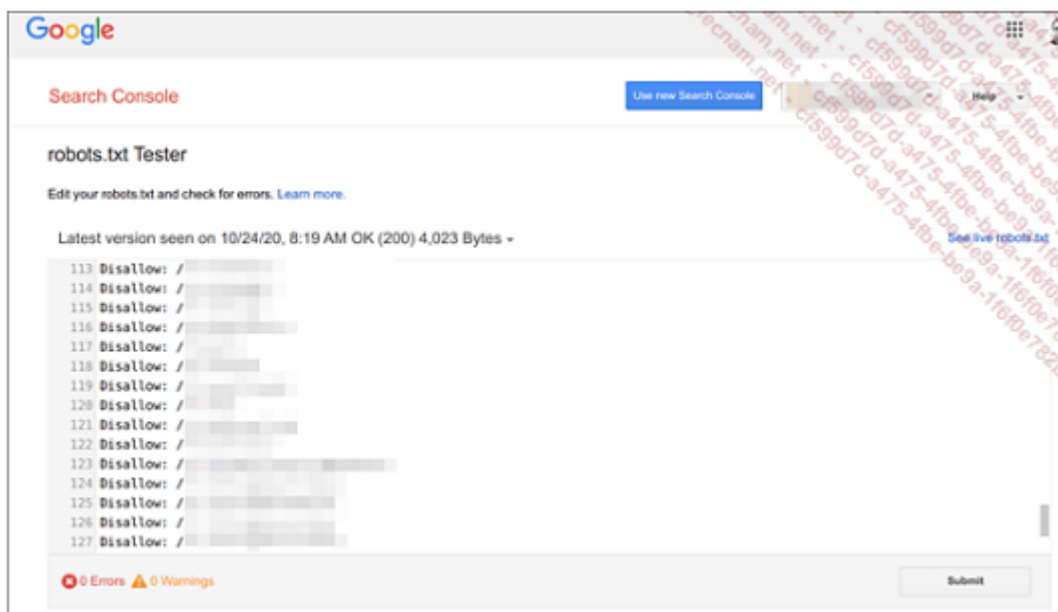
Le fichier robots.txt est-il présent ?

Il doit se trouver à la racine du site. Il est même possible de créer un fichier robots.txt par sous-domaine (même vide) afin d'éviter des erreurs 404, les robots de moteurs de recherche interrogeant les serveurs sur ce point.

Le fichier robots.txt comporte-t-il des erreurs ?

Pour vous en assurer, vous pouvez utiliser les outils suivants :

- Documentation officielle sur l'outil de test de Google Search Console
: <https://support.google.com/webmasters/answer/6062598?hl=fr>
- Outil de test de Google Search Console
: <https://www.google.com/webmasters/tools/robots-testing-tool>
- Outil de test du site technicalseo.com
: <https://technicalseo.com/tools/robots-txt/>



Vous pourrez notamment vous assurer que le fichier robots.txt ne comporte pas de directives 'noindex', celles-ci n'étant plus prises en compte par Google depuis 2019.

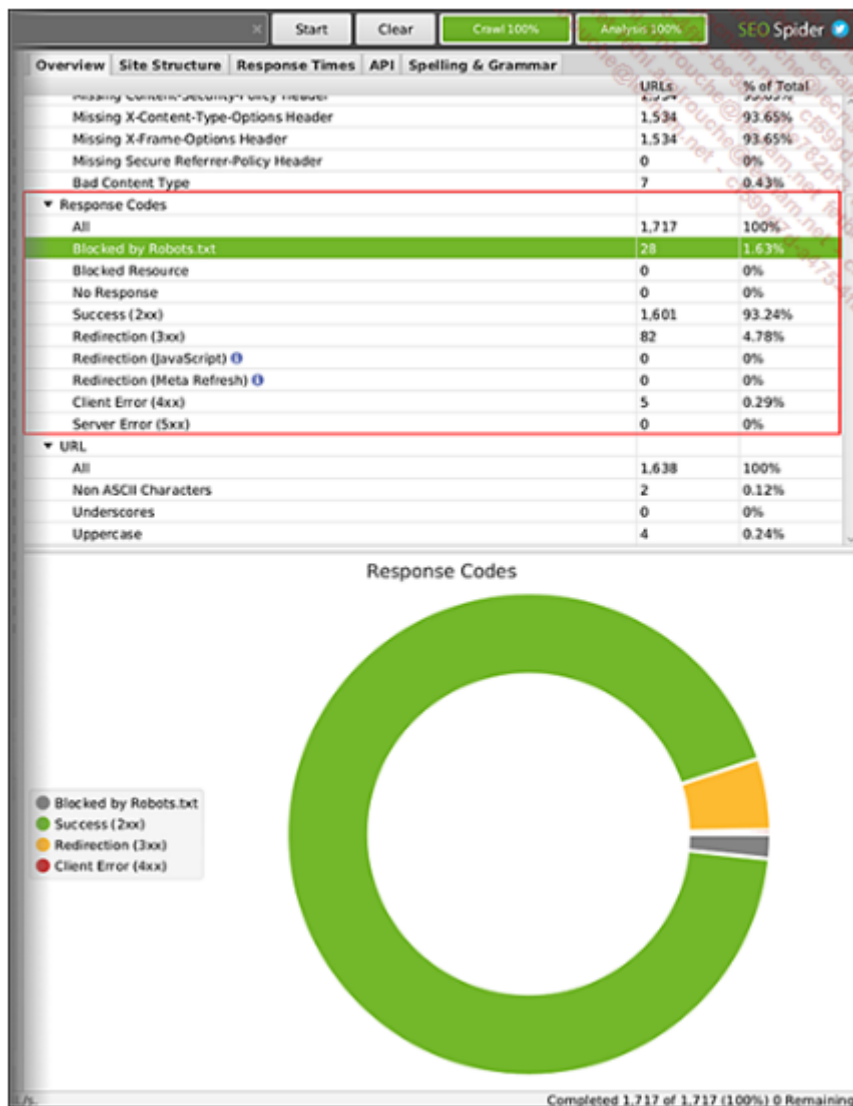
Passage à un site en production C'est un écueil qui peut se produire lors du passage d'un site en développement vers sa version en production. Lors de la phase de développement, les développeurs créent un fichier robots.txt qui bloque le crawl à tous les robots qui pourraient rencontrer le site. Normalement, un site en développement n'est pas accessible depuis l'extérieur, mais sait-on jamais. Il s'agit d'une protection supplémentaire. Le problème, c'est lorsque le fichier robots.txt n'est pas remplacé lors du transfert vers la version de production du site. Le site n'est tout bonnement pas crawlable.

Éléments bloqués par le fichier robots.txt

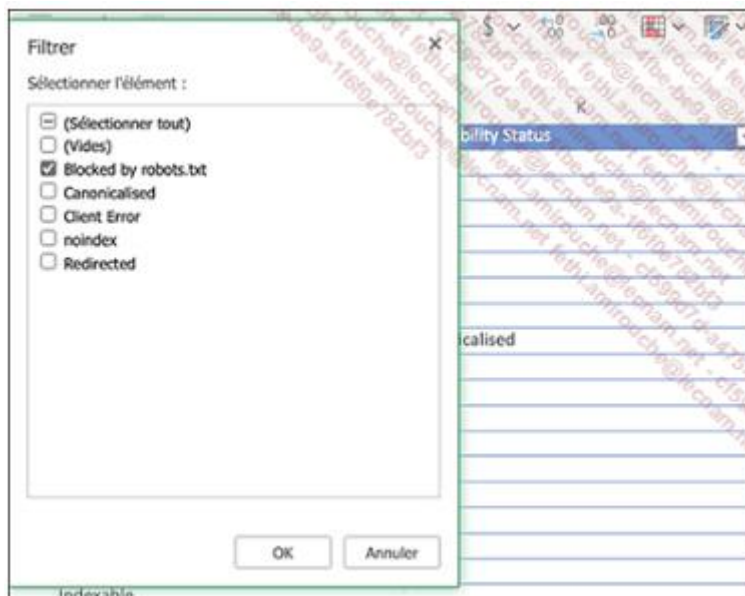
C'est la partie qui nous intéresse le plus. Vous avez à votre disposition deux outils pour vérifier les éléments bloqués : Screaming Frog et Excel.

Analyse avec Screaming Frog

Rendez-vous dans l'onglet **Overview** et choisissez **Response Codes - Blocked by robots.txt**. Vous obtiendrez ici l'ensemble des ressources concernées (pages web, images, etc.).



Avec Excel, vous pouvez filtrer la colonne **Indexability Status** et choisir l'élément **Blocked by robots.txt**.



4. Statut des éléments crawlés

Cette étape s'intéresse aux ressources (pages web, images, CSS, JS, SWF, PDF, etc.) qui ont été crawlées par Screaming Frog, qu'elles soient indexables ou non. Il va de soi que selon la configuration du logiciel les éléments bloqués par le fichier robots.txt peuvent apparaître ou pas dans les résultats. Les préconisations données dans la partie sur la réalisation d'un crawl vous aideront à faire ressortir l'ensemble des éléments du site.

Les différents statuts d'indexabilité des pages web

Pages indexables :

- Pages avec code HTTP 200.
- Pages non bloquées au crawl par le fichier robots.txt.
- Pages non bloquées à l'indexation par une directive noindex (balise meta - robots, en-tête HTTP), les directives dans le robots.tx sont exclues, car elles ne sont plus prises en compte par Google depuis 2019.
- Pages n'ayant pas de redirection (301/302/Meta Refresh).
- Pages ayant une balise *canonical* autoréférente.

Pages non indexables :

- Pages avec code 200 possibles, mais bloquées par robots.txt ou par une directive noindex ou ayant une redirection par balise Meta Refresh.
- Pages avec codes HTTP 301/302/307 (redirections).
- Pages avec codes HTTP 4xx/5xx (erreurs).

- Pages canonisées c'est-à-dire ayant une balise *canonical* vers une autre page.

Analyse du statut d'indexabilité des ressources du site

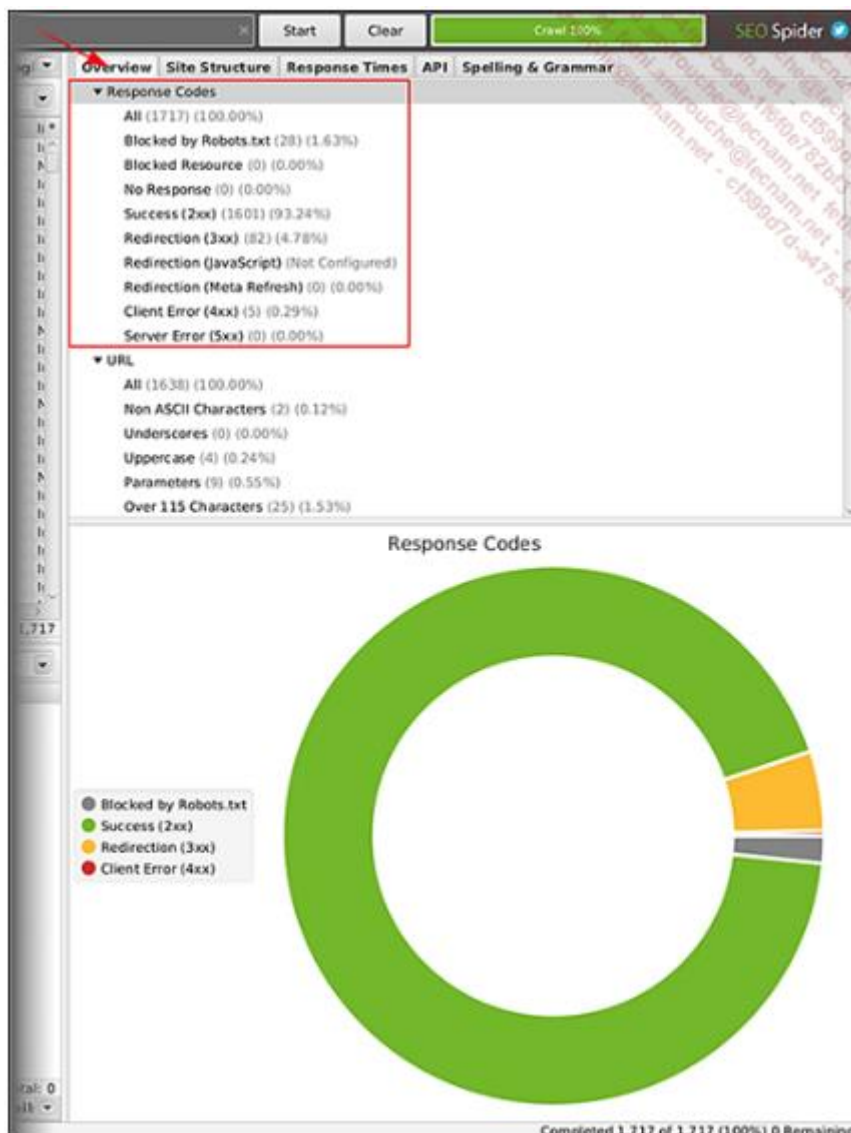
Nous allons faire remonter ici quels sont les statuts rencontrés par Screaming Frog en crawlant le site, à condition que vous ayez réalisé le crawl en demandant au logiciel de récupérer les informations sur tous les éléments du site (les informations qu'il vous fournira concerneront l'ensemble de ces éléments).

Pour faire une analyse uniquement sur les pages, deux solutions sont possibles : crawler de nouveau le site en ne gardant que les pages web, ou utiliser Excel et le fichier nomdedomaine.tld-PAGES.xlsx qui contient toutes ces informations. Voyons ces éléments en détail.

Présenter l'état de tous les éléments du site :

Dans Screaming Frog, allez dans le panneau de droite et cliquez sur l'onglet **Overview**.

Cherchez la rubrique **Response Codes**.



Vous obtenez le détail des codes serveurs rencontrés par Screaming Frog (200, 301, Blocked by robots.txt, etc.). Comme toutes les données du panneau **Overview**, ces informations se trouvent dans le rapport Crawl Overview déjà mentionné.

Vous retrouvez également les éléments bloqués par le fichier robots.txt mentionnés précédemment.

L'intérêt de ces analyses porte essentiellement sur les pages web. Nous allons donc les analyser en profondeur.

Existence d'éléments en HTTP et HTTPS

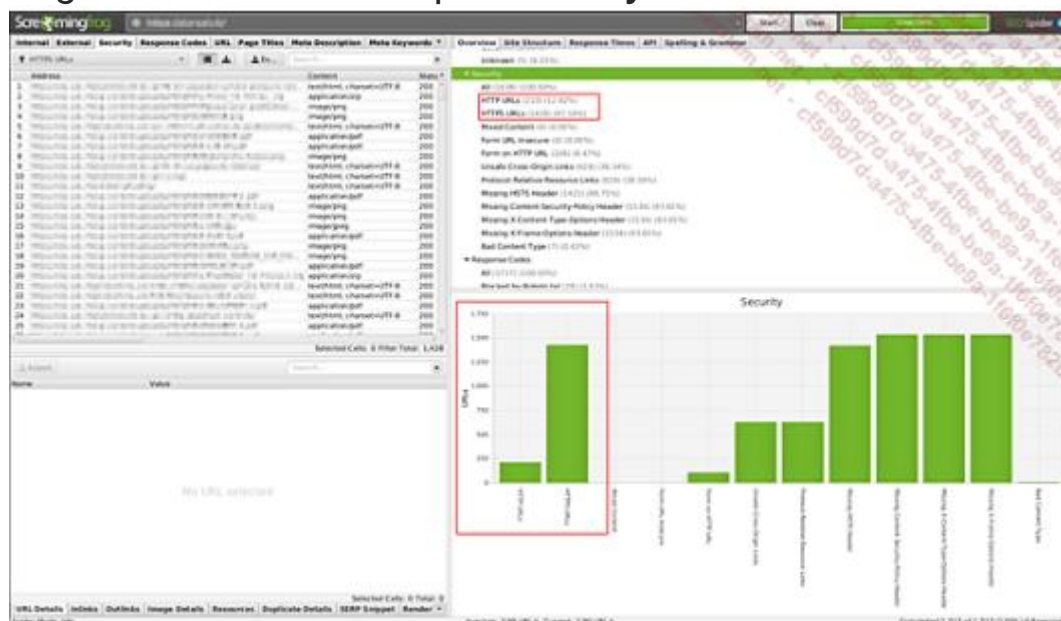
Il est important ici de découvrir s'il existe des ressources en HTTP sur le site, conjointement avec des ressources en HTTPS, et notamment des pages web. Cette situation peut se produire en

raison d'un mauvais paramétrage du serveur web qui laisse accessible des pages dans les deux protocoles. Les problèmes de cette coexistence sont les suivants :

- Les pages en HTTP sont dépréciées par rapport à leurs homologues en HTTPS, **Google favorisant les sites en HTTPS**. En 2018, lors du passage au tout HTTPS lancé par Google, les pages des sites HTTP ont été reléguées en mauvaise position par rapport au site ayant fait l'effort de passer au HTTPS.
- Il existe un possible risque de contenu dupliqué si les pages coexistent dans les deux protocoles, et en l'absence d'indications complémentaires (balises *canonical*, redirections, noindex).
- Même si les pages sont redirigées (redirection 301), pourquoi les maintenir en vie ou permettre qu'elles soient crawlées ? Elles gaspillent du budget crawl.

Analyse rapide avec Screaming Frog

Dans Screaming Frog, rendez-vous dans le panneau de gauche, onglet **Overview** - Rubrique **Security**.



Le logiciel propose des graphiques qui sont très utiles pour agrémenter vos rapports. Intéressons-nous maintenant aux pages HTML.

Analyse des pages HTTP vs HTTPS : vision globale

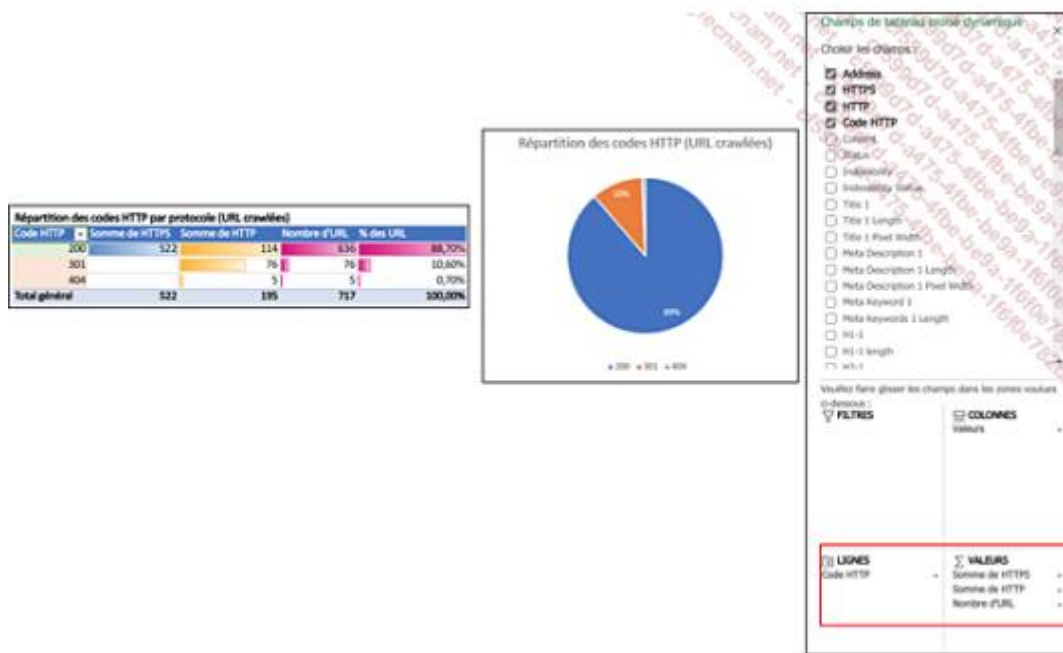
Code HTTP	Somme de HTTPS	Somme de HTTP	Nombre d'URL	% des URL
200	522	114	636	88,70%
301		76	76	10,60%
404		5	5	0,70%
Total général	522	195	717	100,00%

Pour la colonne HTTPS, mettez 1 dans chaque cellule correspondant à une URL commençant par « https ». Pour cela, vous pouvez utiliser les filtres de la colonne **Address**.

[illegible]

Vous pouvez maintenant créer votre tableau croisé dynamique en utilisant les paramètres suivants :

- Tableau de référence : URL
- Feuille sur laquelle placer le tableau : Stats
- Section LIGNES : Status Code (renommé en Code HTTP)
- Section VALEURS : Somme de HTTPS, somme de HTTP, Nombre de Address (ici renommé en Nombre d'URL)



Dans notre cas, il n'y a pas eu d'erreurs 5xx, mais celles-ci seraient à inclure dans la liste au même titre que les codes 301, 302 ou encore 404.

Vous remarquerez que dans le tableau j'ai choisi d'utiliser une couleur de cellule différente (ici du orange) pour les pages en erreur pour les différencier des pages avec un code 200 (ici en vert). Je compte également les pages avec une redirection 301 en erreur. Bien qu'il ne s'agisse pas à proprement parler d'erreurs ou d'un véritable problème, ces pages ne remontent pas avec un code 200 et engendrent des étapes supplémentaires dans le crawl.

Maintenant que nous avons obtenu une vision globale de la situation, intéressons-nous aux pages HTTP puis aux pages HTTPS.

Analyse des pages HTTP

L'intérêt de cette analyse est de vérifier si ces pages sont indexables ou non et de proposer des solutions. Nous allons ici réaliser plusieurs tris avec Excel.

Filtrez la colonne **HTTP** nouvellement créée afin de n'afficher que les cellules contenant des 1. Le tableau va désormais faire ressortir uniquement les pages HTTP.

Analyse des pages HTTP indexables

Dans la colonne **Indexability**, nous allons choisir les pages indexables, si elles sont présentes bien sûr. Dans ce cas, la recommandation est simple : ces pages doivent disparaître ! Concrètement, plusieurs scénarios sont envisageables.

- **Ces pages présentent un équivalent HTTPS** : le plus simple est de mettre en place une politique HSTS sur le serveur pour qu'il redirige automatiquement les pages HTTP vers leur pendant HTTPS. Le HSTS (*HTTP Strict Transport Security*) est un mécanisme de sécurité obligeant un User-Agent à utiliser une connexion sécurisée pour se connecter au serveur. Une autre solution consiste à créer une redirection 301 vers la page HTTPS, mais c'est moins intéressant, car ce type de redirection doit être réalisé à la volée et rajoute une étape supplémentaire.
- **Si ces pages n'ont plus aucune raison d'exister** : elles n'ont pas été remplacées par leur équivalent HTTPS, il s'agit par exemple d'un reliquat d'une ancienne version du site. Dans ce cas, il est nécessaire de les désindexer.

Analyse des pages HTTP non indexables

Dans la colonne **Indexability**, nous allons choisir les pages non indexables, si elles sont présentes. Plusieurs solutions s'offrent également à vous :

- **Pour les pages ayant une directive *noindex*** : le **noindex** est une option intéressante, car elle empêche ces pages d'être présentes dans les SERP. Cependant elles peuvent toujours être découvertes par les moteurs de recherche, ce qui

consomme du budget crawl. On peut envisager un blocage par le fichier robots.txt selon le cas de figure.

- **Pour les pages en redirection 301** : vérifier si les pages de renvoi sont les pages HTTPS. La solution idéale reste ici la mise en place du HSTS.
- **Pour les pages en erreur 4xx** : différents scénarios peuvent se présenter et il est difficile de donner une réponse générique. Les solutions suivantes peuvent être envisagées selon le contexte : mise en place du HSTS si la page HTTPS existe, création d'une page 404 personnalisée si ce n'est pas le cas, suppression des liens entrants sur la page, ou encore redirection vers la page HTTPS si elle existe.
- **Pour les pages canonisées** : si elles renvoient vers leur équivalent HTTPS, une redirection 301 est grandement préconisée. Si elles renvoient vers d'autres pages HTTP, il faudra traiter ce problème au cas par cas.
- **Les pages avec redirection automatique via une balise Meta Refresh** : ici aussi, une réponse générique sera difficile. Une page avec une balise Meta Refresh n'est pas indexable, car elle ne sert que d'étape vers une page finale. Si cette page finale est la version HTTPS de la page, on peut envisager la mise en place du HSTS ou une redirection 301 au pire. Si elle renvoie vers une page en HTTP, il s'agit d'une erreur plus importante qui devra être traitée au cas par cas.

Analyses des pages HTTPS

Nous allons utiliser une analyse similaire à celle des pages HTTP, mais nous nous focaliserons sur les pages non indexables. L'idée d'un site est d'avoir autant que possible de pages indexables (c'est-à-dire dans l'idéal 100 % des pages crawlées).

Analyse des pages HTTPS non indexables

Dans Excel, filtrez la colonne **HTTPS** nouvellement créée en ne gardant que les cellules contenant le chiffre 1.

Filtrez ensuite la colonne **Indexability** et choisissez l'élément **Non-Indexable**.

Triez ensuite les données selon les différents cas de figure qui se présentent, c'est-à-dire des directives noindex, des pages

canonisées, des redirections 301/302, des pages en erreur ou des redirections Meta Refresh.

Le principe sous-jacent à ces préconisations étant d'éviter de consommer du budget crawl inutilement.

Analyse des pages HTTPS indexables

Ici, nous allons nous intéresser au cas contraire : des pages indexables, mais ne devant pas l'être.

Analyse des pages indexables

Dans le fichier Excel, tout en gardant la colonne **HTTPS** filtrée sur les cellules contenant le chiffre 1, filtrez la colonne **Indexability** et choisissez l'élément **Indexable**.

Ici aussi, il vous faudra opérer au cas par cas. Des URLs contenant des paramètres ressortent-elles ? Certaines catégories de pages ne devraient-elles pas être présentes ?

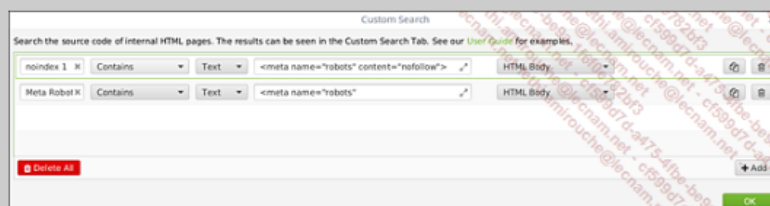
Le principe est d'économiser du budget crawl et d'éviter que des pages ressortent dans les SERP, ce qui est potentiellement le cas avec ce type de pages.

Problèmes avec le sous-domaine www

Il est possible pour un site d'utiliser le sous-domaine www ou d'opter pour un domaine racine. Mais il faudra éviter que des pages soient disponibles avec des URLs comprenant à la fois www et ne le comprenant pas. Cette situation génère en effet du contenu dupliqué.

Reprenez le filtrage mentionné ci-dessus et partez à la recherche de pages mentionnant ou ne mentionnant pas **www** (dans la colonne **Address** du fichier Excel).

Astuce 1 : custom filter pour trouver les pages avec un double noindex. Google est capable de reconnaître les pages ayant des balises `meta robots noindex` situées ailleurs que dans l'en-tête HTML `<head>`, notamment dans le `<body>`. Il s'agit bien de la balise `meta robots`, et non de texte écrit dans un paragraphe (entre des balises `<p>`). Ceci peut entraîner une désindexation de la page dans certaines conditions. De même, si une page possède à la fois des balises `meta robots index` et `noindex`, Google retiendra cette dernière. Pour plus d'informations, vous pouvez consulter la page : <https://www.webrankinfo.com/dossiers/indexation/balise-meta-robots-noindex>. Heureusement, il est possible d'utiliser Screaming Frog pour détecter ce genre d'erreurs. Pour cela, vous pouvez utiliser les fonctions de recherche personnalisée se trouvant dans le menu **Configuration - Custom - Search**. Vous créez deux règles de recherches, l'une portant sur le texte `<meta name="robots" content="nofollow">` et la deuxième sur `<meta name="robots">`. La zone d'analyse à sélectionner sera HTML Body.



Astuce 2 : vérifier si la page d'accueil est indexable. Cela peut prêter à sourire, mais c'est un point primordial dont il s'agit de s'assurer : la ou les pages d'accueil du site sont-elles indexables ? On imagine aisément les dégâts de ce type de mauvaise configuration sur le référencement d'un site.

5. Trouver les pages non indexables (deuxième méthode)

Cette partie est similaire à la précédente, mais utilise une autre approche en se focalisant sur l'étude des pages non indexables de manière générale. Avant d'entrer dans le vif du sujet, voici quelques éléments théoriques :

Il existe plusieurs moyens techniques pour bloquer l'indexation des pages d'un site, parmi lesquels :

- **Protéger des répertoires** avec un identifiant et mot de passe (pour un serveur Apache, il s'agit de configurer le fichier `.htaccess`). C'est une approche radicale, mais qui est utile pour les sites en recette/staging (en cours de migration ou de construction), mais aussi pour utiliser des outils en ligne de manière sécurisée.

- **Configurer le fichier robots.txt** pour bloquer l'accès de manière granulaire (à certains robots, aux images, à certains répertoires, à certaines pages, etc.).
- En donnant des directives directement dans les pages web avec la balise `meta robots`.
- En donnant des **directives dans les en-têtes HTTP** (pour les formats tels que le PDF, XML, fichiers Word, Excel ou encore PowerPoint).

Comme nous l'avons déjà vu, on peut vouloir **empêcher l'indexation** de pages ou d'autres ressources pour plusieurs raisons. Tout d'abord pour **préserver la confidentialité des données**, comme c'est le cas des sites en cours de développement ou de migration. Ensuite, pour **économiser du budget crawl** en excluant du crawl des pages sans aucun intérêt pour l'utilisateur (ex. : accès à l'interface d'administration d'un CMS). Et enfin pour faire en sorte que des **pages de moindre importance ne ressortent pas** dans les SERP, telles que des pages à faible contenu ou au contenu dupliqué (dans un site e-commerce, cela concerne un même produit décliné en plusieurs tailles ou couleur par exemple).

Dans cette analyse, nous allons nous pencher sur les éléments suivants :

- Pages ou ressources ne pouvant pas être crawlées du fait du fichier robots.txt.
- Pages non indexables du fait de nombreux facteurs (noindex, canonisées, erreurs, etc.).

Analyser les éléments bloqués par le fichier robots.txt

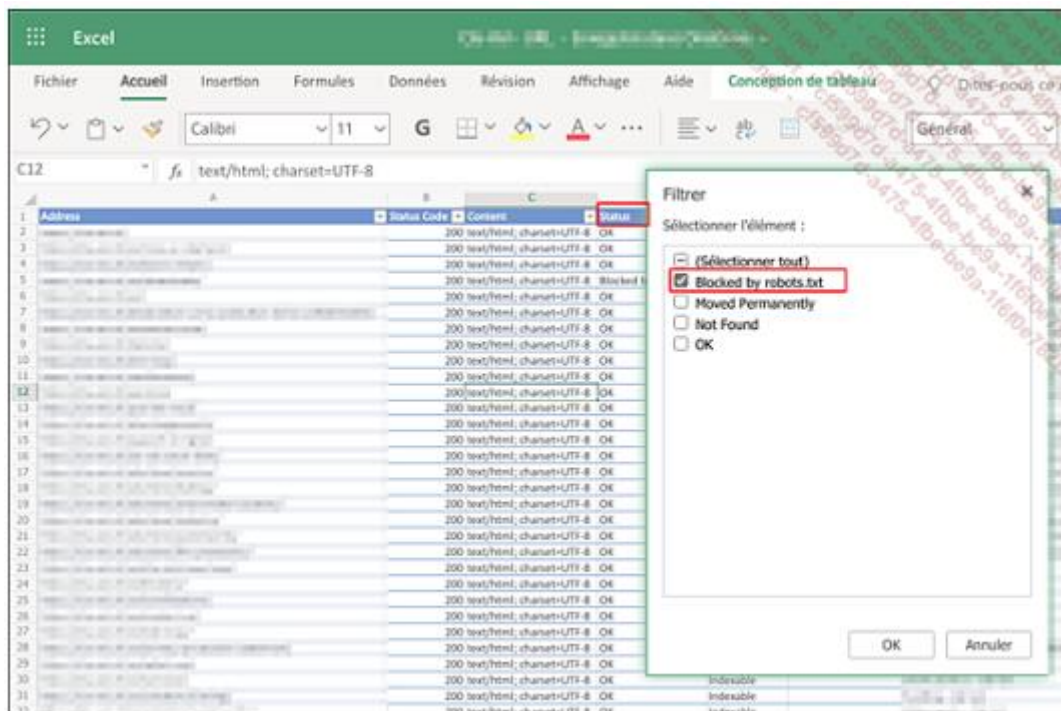
Analyser les pages dans Excel

Ici, nous allons nous baser sur une analyse des pages web uniquement en utilisant le fichier d'export (nomdedomaine.tld-PAGES.xlsx). Il sera possible, ce que nous verrons dans la section suivante, de réaliser cette opération sur l'ensemble des éléments du site.

Filtrez le fichier nomdedomaine.tld-PAGES.xlsx sur la colonne **Status** et choisissez l'élément **Blocked by robots.txt**, si

cette entrée est disponible. Si elle n'apparaît pas dans la liste, c'est qu'aucune page n'est bloquée ou que les parties du site qui sont mentionnées dans le fichier robots.txt n'ont pas de liens entrants (ex. : la partie d'administration d'un site). Cette situation devrait vous alerter.

L'analyse avec Screaming Frog ci-après montre les autres éléments potentiellement bloqués, mais qui ne sont pas des pages web.



Si des pages apparaissent dans cette section, ce sera à vous de juger si elles doivent effectivement être bloquées ou pas, selon votre connaissance du site ; vous pourriez par exemple estimer que des pages de catalogue produits destinés à l'équipe commerciale, et non aux clients potentiels, devraient être exclues.

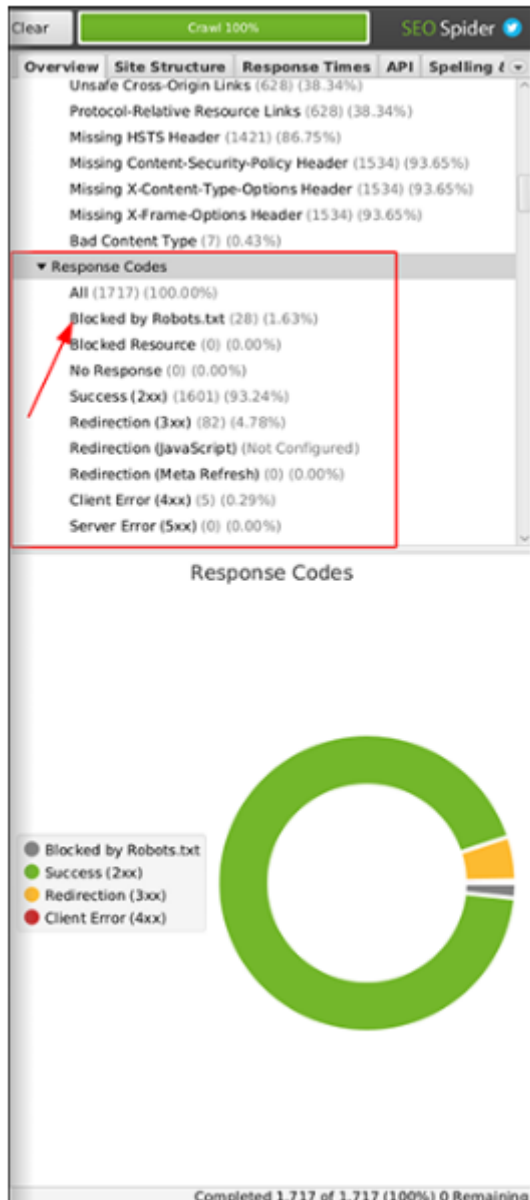
Analyser les ressources avec Screaming Frog

Complétez votre analyse par les données de Screaming Frog afin d'avoir un aperçu global de la situation. En effet, l'analyse ci-dessus s'intéressait uniquement aux pages web puisque ce sont elles que nous avons exportées. Screaming Frog possède toutes les informations relatives aux autres types de fichiers.

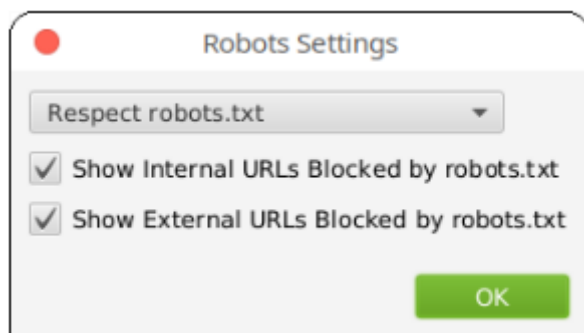
Dans le panneau de droite (Overview), cherchez le champ **Response Codes** et cliquez sur **Blocked by robots.txt**. La fenêtre principale va

alors basculer sur l'onglet correspondant et filtrer les données sur le statut **Blocked by robots.txt**.

Vous trouverez ici l'ensemble des fichiers bloqués par le fichier robots.txt : pages web, images, fichiers JavaScript, etc.



Vous ne voyez toujours rien ? Soit le fichier robots.txt est absent ou vide, soit vous avez mal configuré Screaming Frog avant le crawl. Si possible relancez un crawl du site avec l'option **Respect robots.txt** et en cochant les deux cases **Show Internal/External URLs Blocked by robots.txt**.

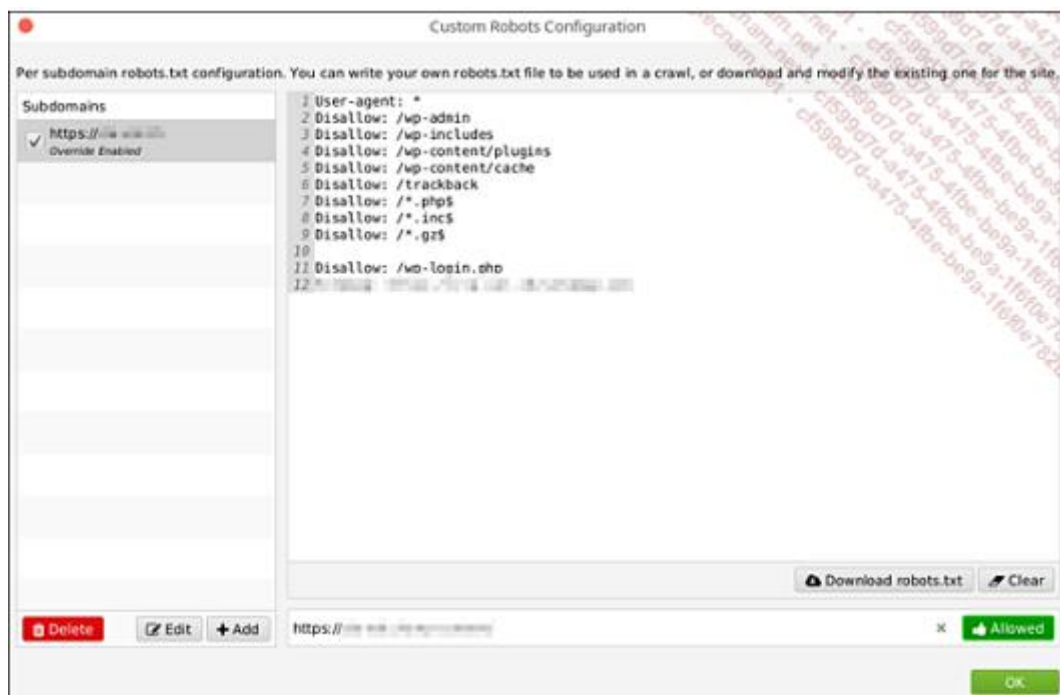


Ajouter un fichier robots.txt à chaque sous-domaine

Si le site possède des sous-domaines, il est nécessaire de rajouter un fichier robots.txt à la racine de chacun d'eux, même si vous les laissez vides. En effet, le Googlebot va aller à la recherche de fichiers robots.txt sur les sous-domaines, et s'il n'en rencontre pas, cela générera une erreur 404.

Screaming Frog offre également la possibilité de **tester votre fichier robots.txt à la volée**. Cette option se trouve dans le menu **Configuration - Robots.txt - Custom**.

Il vous suffit d'ajouter le nom du site (en sous-domaines). Le logiciel va récupérer le contenu du fichier robots.txt qui se trouve à la racine du répertoire et vous permettre de tester en temps réel des URLs et de voir si elles sont bloquées ou pas.



Frog : <https://www.screamingfrog.co.uk/robots-txt-tester/>

Méthodologie

Tous les sites sont différents. Selon le nombre de pages ou encore le type de blocages que vous découvrirez, il vous sera utile de faire soit une analyse générale soit une analyse dans le détail.

Pour une analyse générale portant sur un petit site, utilisez Excel :

Dans la colonne **Indexability** du fichier nomdedomaine.tld-PAGES.xlsx, filtrez le contenu **Non-Indexable**.

Vous verrez apparaître les différentes pages ainsi que les raisons de leur non-indexabilité dans la colonne suivante **Indexability Status**. Vous pourrez alors vérifier rapidement si ces pages sont en erreur ou pas

Pour un site plus important, optez pour Screaming Frog et les informations ci-après.

[illegible]

Vérifiez bien que la homepage (ou les homepages) sont indexables. Il peut arriver que ces pages ne soient pas indexables. Ce peut être notamment le cas lors de migrations ou d'ajout d'une nouvelle langue à un site existant, la balise *canonical*/renvoyant vers la page initiale du site par exemple.

Analyser les pages en erreur 301/302/307, 4xx, 5xx, Meta Refresh

Comme vous pouvez le constater, j'ai mis les codes HTTP 4xx (404, 410), 5xx et 301/302 au même niveau. S'il est vrai que les erreurs 404 sont effectivement des erreurs (elles peuvent cependant être utilisées pour désindexer des pages), les codes 301

peuvent également être considérés comme (légèrement) problématiques. Ils rajoutent des niveaux de profondeur et consomment du budget crawl de manière inutile. Les choses sont à identifier au cas par cas. On peut également ajouter à la liste des redirections, les codes HTTP 307. Pour rappel, ils correspondent aux redirections faites par une page en HTTP avec un renvoi automatique vers sa version HTTPS si vous avez activé l'option **Respect HSTS Policy** dans le menu **Configuration - Spider - Advanced**.

Tout comme pour la recherche des erreurs :

Utilisez l'onglet **Overview** de la colonne de droite de Screaming Frog.

Allez à la rubrique **Response Codes**.

Et faites votre choix entre les différents éléments proposés.

Comme précédemment, toutes ces données peuvent être exportées séparément pour être fournies comme documents annexes à votre client. Vous pourrez retraiter ces données avec Excel en filtrant la colonne **Content Type** sur l'élément **text/HTML ; charset=UTF-8** qui fait référence aux pages web ou encore sur **application/pdf** pour les fichiers PDF.

Cette manipulation peut également être réalisée dans Excel avec le fichier nomdedomaine.tld-PAGES.xlsx (et ne concerne donc que les pages web), en filtrant les colonnes **Status Code** et **Meta Refresh 1**.

Quelles corrections apporter ?

Il s'agit d'un autre travail d'analyse qui va se présenter ici. En ce qui concerne les erreurs 404 (soft et hard) par exemple, vérifiez que les pages sont en erreur 404 ou bien si ce sont les liens pointant vers ces URLs qui sont défectueuses :

- Exportez tous les liens internes (**Bulk Export - All Inlinks**).
- Filtrez par **Code Status** et choisissez **404**.
- Vérifiez les pages et les liens (colonne **Destination**).

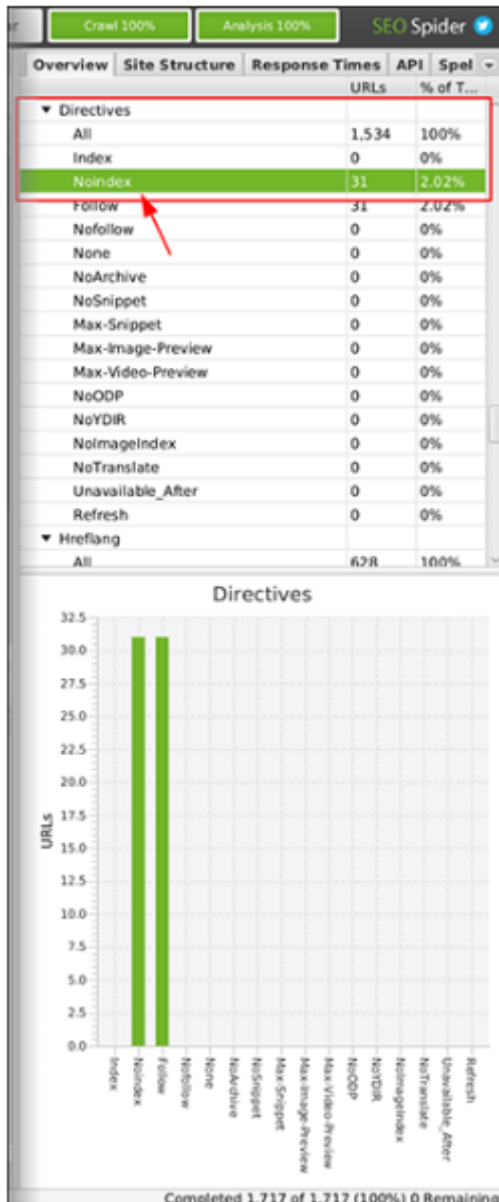
Trouver les pages avec balises noindex

Vous pouvez tout aussi bien réaliser cette manipulation dans Screaming Frog que dans Excel. Screaming Frog fournit les renseignements pour toutes les ressources du site qui ont été

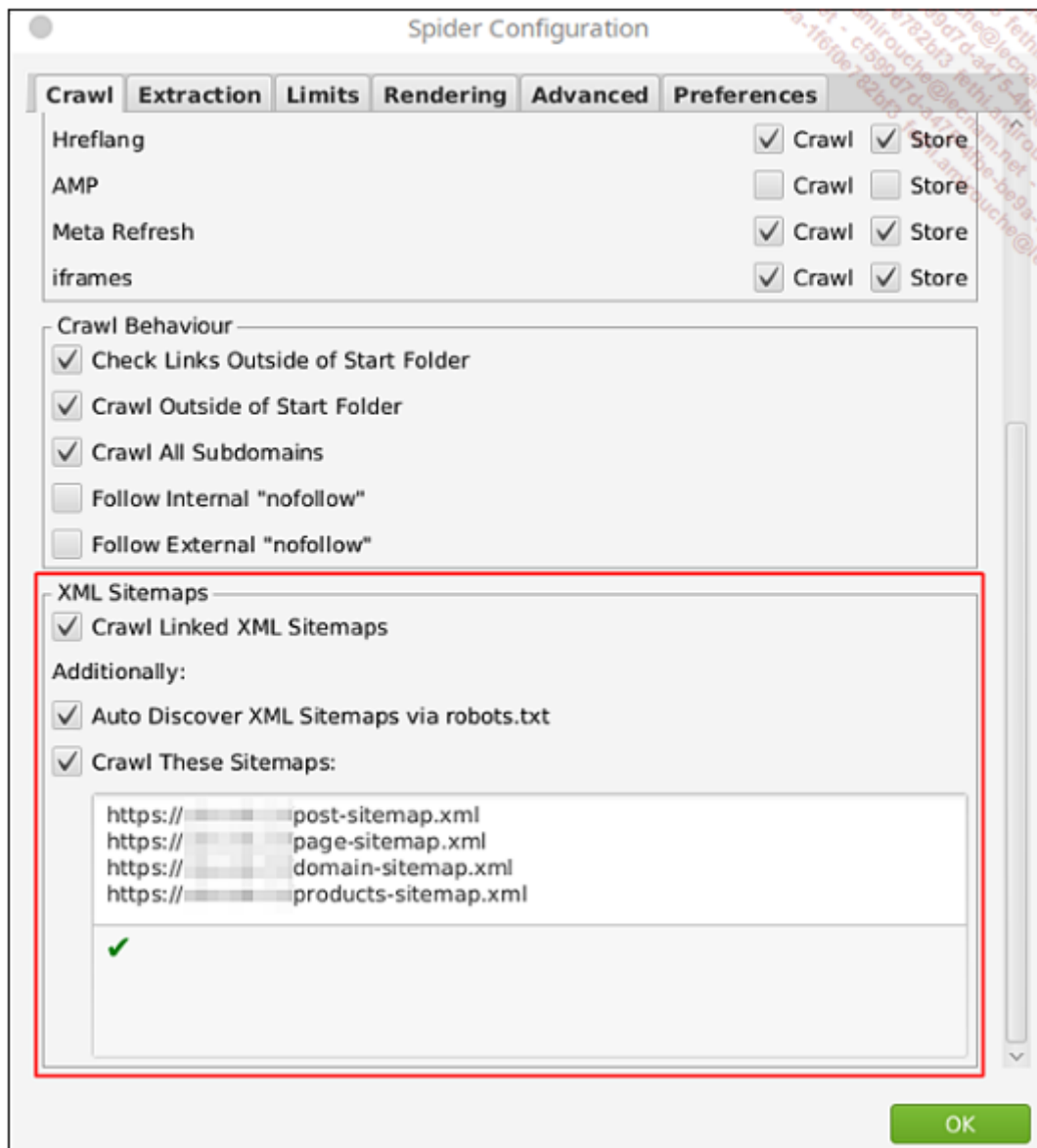
crawlées, tandis qu'avec Excel et le fichier nomdedomaine.tld-PAGES.xlsx, vous vous concentrerez uniquement sur les pages.

Trouver les éléments avec des directives nofollow dans Screaming Frog

Dans l'onglet de droite **Overview**, cliquez sur **Directives** et la sous-catégorie **Noindex**.



Le résultat est le suivant :



Le prérequis pour cette analyse est d'avoir bien configuré les sitemaps dans le menu **Configuration - Spider** - onglet **Crawl**.

La prochaine étape consiste à demander à Screaming Frog de générer une analyse du crawl portant sur les sitemaps. En vous rendant dans l'onglet **Sitemaps** vous constaterez qu'il est vide, à l'inverse de la plupart des autres onglets. Le message suivant apparaît : « You need to perform Crawl Analysis in order to populate this filter ».

Allez dans le menu **Crawl Analysis - Start**. La boîte de dialogue suivante apparaît.

Crawl Analysis Configuration

Some data can only be calculated by analysing the crawl when it is paused or ...

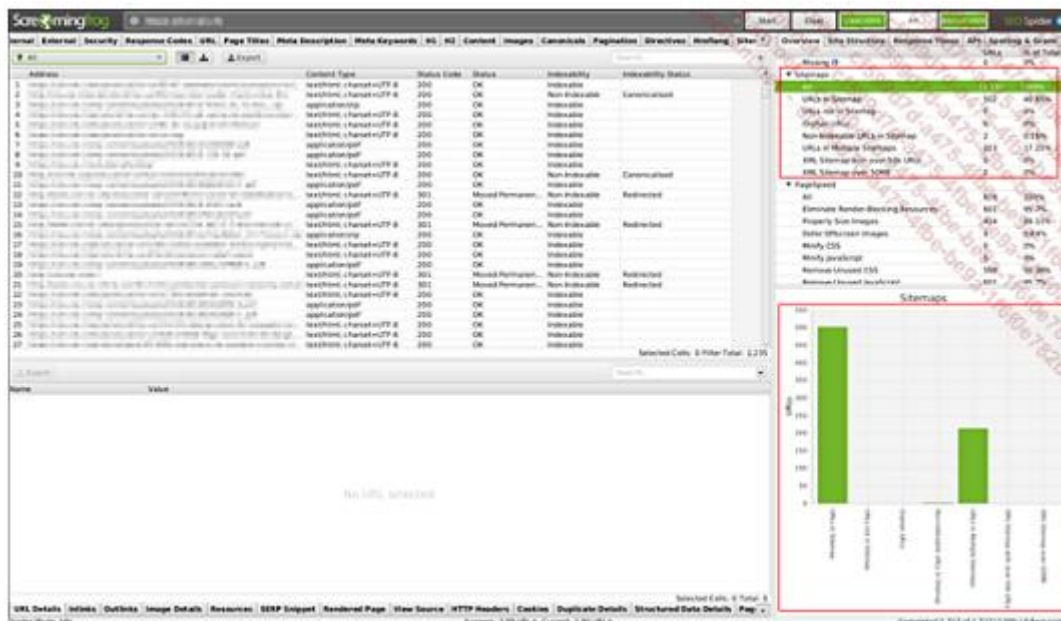
- ☐ **Link Score**
Assigns a Link Score to all internal URLs
- ☐ **Content**
Near Duplicates
- ☐ **Pagination**
Unlinked Pagination URLs, Pagination Loop
- ☐ **Hreflang**
Unlinked hreflang URLs, Missing
- ☐ **AMP**
Non-200 Response, Missing <html amp> Tag
- ☒ **Sitemaps**
URLs in Sitemap, URLs not in Sitemap, Orphan URLs, Non-Indexable URLs in Sitemap, URLs in Multiple Si...
- ☐ **Analytics**
Orphan URLs
- ☐ **Search Console**
Orphan URLs

☐ Auto-analyse at End of Crawl

OK

Sélectionnez la case **Sitemaps** et cliquez sur le bouton **OK**. Screaming Frog remplit alors la liste des URLs et des potentiels problèmes liés aux sitemaps.

Vous pourrez visualiser ces informations de manière graphique comme dans chaque analyse proposée par Screaming Frog :



URLs In Sitemap

Liste l'ensemble des URLs se trouvant dans les sitemaps. Cette section ne devrait contenir que les URLs des pages indexables et canoniques des pages importantes du site.

URL Not In Sitemap

Comme son nom l'indique, il s'agit des URLs non présentes dans le sitemap, mais que Screaming Frog a détectées au cours du crawl. Si tout va bien, ces URLs sont celles des pages de faible importance, et donc il est normal qu'elles ne figurent pas dans cette rubrique. Ou alors, ces URLs sont réellement absentes des sitemaps et devraient y figurer. Attention toutefois : ce filtre ne prend en compte que les URLs reconnues comme indexables (pas de noindex ou de robots.txt bloquant leur accès).

Orphan URLs

Les **URLs orphelines** sont des URLs pouvant être découvertes par plusieurs moyens (fichiers sitemaps, Google Search Console et Google Analytics). Elles apparaissent dans ces données, mais ne peuvent être trouvées par un robot, car elles ne bénéficient d'aucun lien entrant sur le site. Elles montrent en quelque sorte des défauts de conception du site. Mais comment se fait-il que ces URLs apparaissent dans les données de Google Analytics ou de Google

Search Console si on ne peut pas les trouver ? Réponse : grâce aux fichiers sitemaps et aux backlinks.

Les fichiers sitemaps étant normalement générés de manière automatique par les CMS, ils peuvent faire ressortir des URLs isolées qui se retrouvent indexées et apparaissent dans les résultats de recherche. Des internautes cliquent dessus et les données de navigation sur ces pages se retrouvent dans Google Analytics et Google Search Console.

Le problème c'est que même si ces URLs sont indexées, elles ne bénéficient pas de performances optimales du fait de l'absence de liens internes entrants (le PageRank ne circule pas). À terme, elles risquent de voir leur score SEO décliner et de perdre des positions. Il est possible également que ces pages doivent disparaître du sitemap car elles sont sans intérêt.

Si vous souhaitez aller plus loin sur ce sujet, vous trouverez une analyse plus complète dans la documentation de Screaming Frog : <https://www.screamingfrog.co.uk/find-orphan-page>

Non-Indexable URLs in Sitemap

Comme son nom l'indique, il s'agit des URLs ayant un problème d'indexabilité (erreurs 4xx, 5xx, etc.), mais qui se trouvent dans le sitemap. À supprimer ou à corriger.

URLs In Multiple Sitemaps

Il s'agit de la liste des URLs se trouvant dans plusieurs sitemaps. Cela n'est pas grave en soi, mais pas optimal pour autant.

XML Sitemap With Over 50k URLs

Le nombre maximal d'URLs se trouvant dans un fichier XML est de 50 000. Il faudra donc veiller à alléger le/les fichiers avec les URLs les plus pertinentes. D'autant qu'il est possible d'indiquer jusqu'à 50 000 fichiers XML pour un site, repoussant la limite d'URLs possiblement indexables à 2,5 milliards.

XML Sitemap With Over 50mb

Le poids maximal d'un fichier sitemap est fixé à 50 Mo. Ici apparaissent les fichiers XML dépassant cette limite.

Aller plus loin avec les fichiers sitemaps. Plus d'informations sont disponibles sur le site en charge du respect des normes des sitemaps : <https://www.sitemaps.org/> (interface en français) ou sur la documentation de Google Search Console : <https://developers.google.com/search/docs/advanced/sitemaps/build-sitemap>

Crawler les fichiers sitemaps avec Screaming Frog

Il est tout à fait possible de vérifier le statut des pages envoyées par vos sitemaps et de remonter les erreurs éventuelles.

Pour ce faire, dans Screaming Frog :

Utilisez le mode **List** disponible dans le menu **Mode - List**.

Téléchargez les fichiers sitemaps que vous souhaitez tester (au format .XML) ou récupérez l'URL du sitemap sur le serveur. S'il s'agit d'un fichier sitemap maître faisant référence aux autres sitemaps du site, Screaming Frog sera capable d'analyser les URLs de chacun des fichiers.

Cliquez sur le bouton **Upload** qui vous propose plusieurs options pour choisir le fichier ou pour indiquer l'URL du fichier sitemap. Lancez le crawl des URLs.

Le résultat sera une analyse de crawl comme celle du site complet, mais qui sera basée uniquement sur les URLs fournies. Dans l'image ci-après, l'ensemble des pages analysées donnent des réponses 200, ce qui est un bon point, mais d'autres analyses plus avancées devront être réalisées.



	Address	Content Type	Status Code	Status
1	http://www.screamingfrog.co.uk/	text/html; charset=UTF-8	200	OK
2	http://www.screamingfrog.co.uk/	text/html; charset=UTF-8	200	OK
3	http://www.screamingfrog.co.uk/	text/html; charset=UTF-8	200	OK

7. Liens de pages avec attributs nofollow, ugc ou sponsored

Un peu d'histoire

Souvenez-vous : avant, il était possible d'indiquer à Google si l'on souhaitait que son robot suive ou pas les liens de pages : il suffisait d'ajouter l'attribut `rel="nofollow"` dans les balises de lien (`<a>`). Ce qui signifie également que par défaut, les liens étaient considérés comme `nofollow`.

Les règles du jeu ont changé avec une mise à jour de Google entrée en vigueur le 1er mars 2020 : Google indique alors qu'il ne tient plus compte des attributs de liens, ou plutôt qu'il considère l'attribut `nofollow` comme un « indice » (« *hint* » en anglais). et non comme une obligation que Google devra respecter scrupuleusement. Il s'agit ici d'indiquer au moteur de recherche comment vous souhaitez voir le lien traité ; mais, libre à lui de respecter cette indication ou de la traiter d'une autre manière. Deux autres attributs ont fait leur apparition : `ugc` et `sponsored` : l'attribut **ugc** est dédié aux liens générés par des utilisateurs d'un site (sur un forum ou des commentaires de blog par exemple), tandis que l'attribut **sponsored** concerne les liens commerciaux tels que les liens d'affiliation.

Voilà quelques détails sur le sujet : <https://webmasters.googleblog.com/2019/09/evolving-nofollow-new-ways-to-identify.html>

Au final, on ne sait pas ce que cela change vraiment, Google pouvant suivre ou pas les liens en `nofollow`. Il peut donc s'avérer intéressant de vérifier si le site audité comporte des liens avec cet attribut et si les pages visées doivent absolument être exclues de l'indexation.

Trouver les liens en `nofollow` avec Screaming Frog

Il est aisé de savoir si les liens sont en `DoFollow` ou utilisent les autres attributs (`nofollow`, `ugc`, `sponsored`).

Faites une extraction des liens internes par le menu **Bulk Export - All Inlinks**. Dans la colonne intitulée **Follow**, les liens suivis portent la valeur **True**, les autres portent la valeur **False**.

Repérez les pages dont le contenu de la colonne **Follow** est **False**. Vous pouvez également filtrer la colonne précédente (**Status Code**) pour y déceler des erreurs éventuelles.

The screenshot shows an Excel spreadsheet with the following columns: **Status**, **Code**, **Import**, **Action**, **Type**, and **Path**. The data is organized into rows, with some rows highlighted in orange. The 'Action' column contains values like 'True' and 'False', and the 'Type' column contains values like 'Absolute' and 'Relative'. The 'Path' column contains file paths like 'C:\Users\user\AppData\Local\Microsoft\Windows\CurrentVersion\Explorer\RecentDocs\'. The spreadsheet is displayed in a window titled 'Excel' with a standard Windows interface.

Il est également possible de rechercher les attributs de liens `noreferrer`, figurant systématiquement sur les liens externes de WordPress notamment, et qui peuvent être un frein dans le cadre d'une stratégie de *linkbuilding*. Pour cela, vous pouvez utiliser les fonctionnalités *Custom Search* de Screaming Frog abordées dans le chapitre précédent.

Les URLs dupliquées sont courantes et sont souvent liées à des défauts de configurations du CMS qui va proposer un contenu identique pour des taxonomies différentes. Ainsi dans WordPress, sans optimisation, on retrouvera plusieurs pages identiques selon que les données soient filtrées par catégorie, par auteur ou par date. Sur un site e-commerce sous Magento, ce seront des pages produits (PLP/PDP) qui pourront être les mêmes selon que le tri soit fait par catégorie de produits ou par marque.

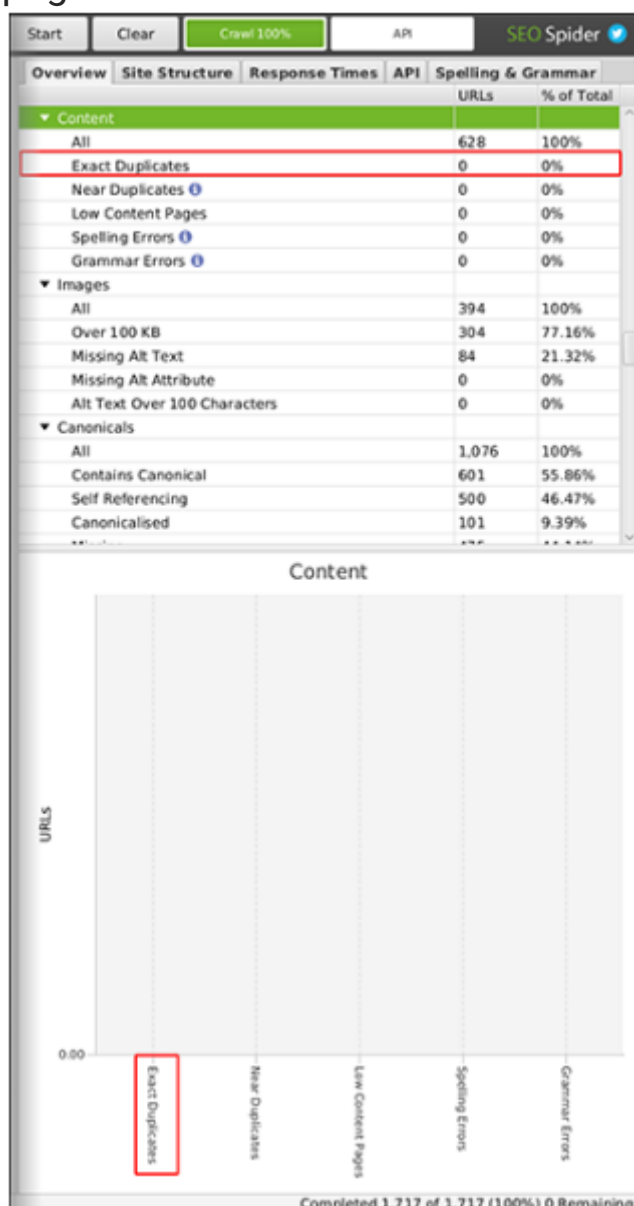
47

strict seront analysées. Mais il est possible de choisir un seuil de tolérance plus ou moins élevé.

Analyse du contenu dupliqué strict

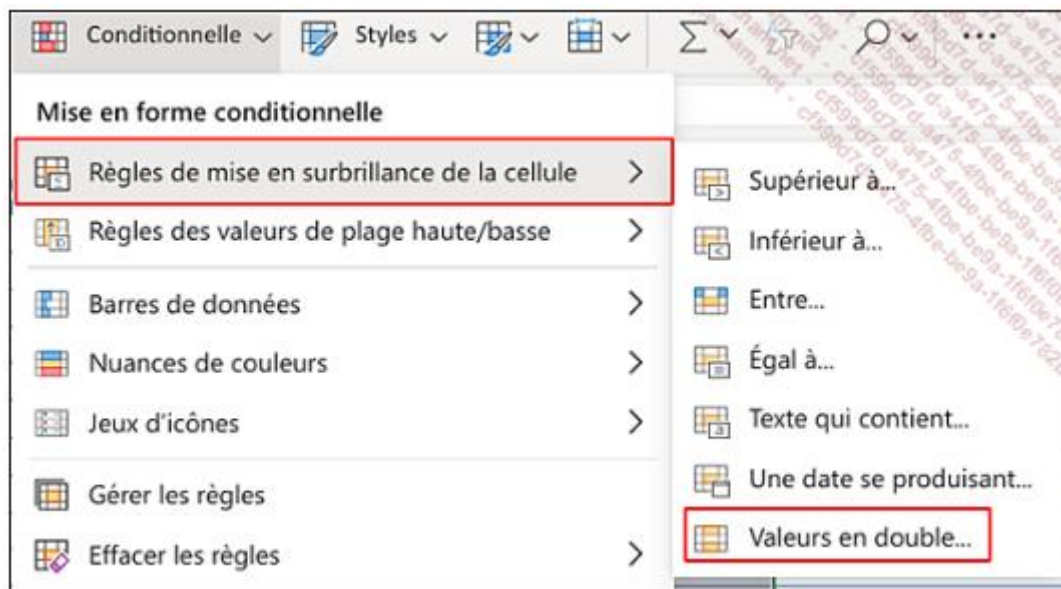
Deux voies d'analyse s'offrent à vous : avec Screaming Frog ou avec Excel.

Dans **Screaming Frog**, allez à l'onglet **Content** pour avoir un aperçu de la colonne **Exact Duplicates** (en français : dupliqués stricts). Vous pouvez trier le menu en haut à gauche pour obtenir la liste des pages incriminées.



Avec **Excel**, utilisez le fichier nomdedomaine.tld-PAGES.xlsx, et dans la colonne **Hash Values**, appliquez une mise en forme

conditionnelle pour mettre en évidence les valeurs en double. Vous verrez apparaître les cellules comportant les mêmes valeurs de Hash. Le Hash de chaque page est obtenu en utilisant un algorithme qui prend en compte l'ensemble du code HTML et le convertit en une chaîne de caractères. Screaming Frog utilise l'algorithme MD5 (il en existe d'autres) et cette chaîne de caractères (ou Hash) est appelée « chaîne de contrôle MD5 ». Ainsi, deux pages au contenu strictement identiques auront la même somme de contrôle.



Analyse du contenu dupliqué proche

Si vous avez paramétré Screaming Frog pour cela (voir la section Analyser le contenu dupliqué du chapitre Préparation de l'audit), il est possible de voir quelles sont les pages qui possèdent un **contenu relativement proche** (*Near Duplicates*). Il s'agit de pages dont le contenu est très similaire et qui risquent d'être considérées comme dupliquées. C'est ainsi le cas sur des sites e-commerce où des pages produits (PLP) décrivent un seul produit, mais plusieurs sous-pages existent pour présenter les déclinaisons (par taille, couleur, public visé, etc.). Un seul caractère sur une page peut affecter la somme de contrôle MD5 donnant les *hash values* qui détectent les pages exactement identiques ; une analyse de ces types de contenus *Near Duplicates* est donc nécessaire.

Réaliser un crawl dédié aux Near Duplicates

Si vous n'avez pas lancé le crawl avec le paramétrage adéquat, vous devrez en réaliser un nouveau pour cette analyse spécifique :

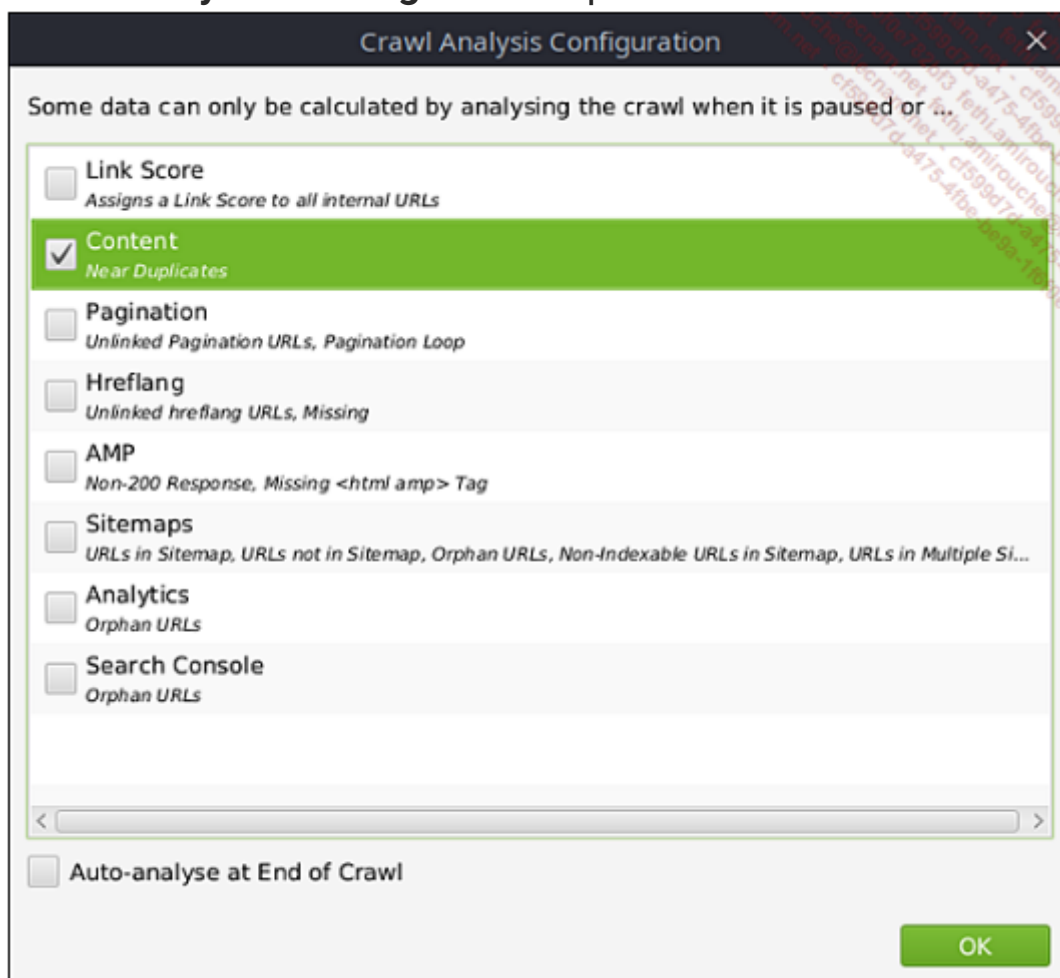
- Reprenez la configuration proposée dans la section relative à la préparation du crawl (section Analyser le contenu dupliqué du chapitre Préparation de l'audit).
- Pour simplifier et gagner du temps, dans la partie **Spider** - onglet **Resources Links**, décochez toutes les options qui pourraient ralentir le crawl (à moins que le site nécessite du JavaScript pour afficher du contenu).

Votre but est de ne crawler que les pages web.

Procédure d'analyse

Vous avez un crawl qui a pris en compte les contenus proches (*Near Duplicates*) ? Bien. À l'inverse du rapport relatif au contenu dupliqué strict, les données d'analyse ne remontent pas de manière automatique. Vous devez demander à Screaming Frog de les générer. Notez que cette manipulation peut être réalisée à la volée, pendant le crawl.

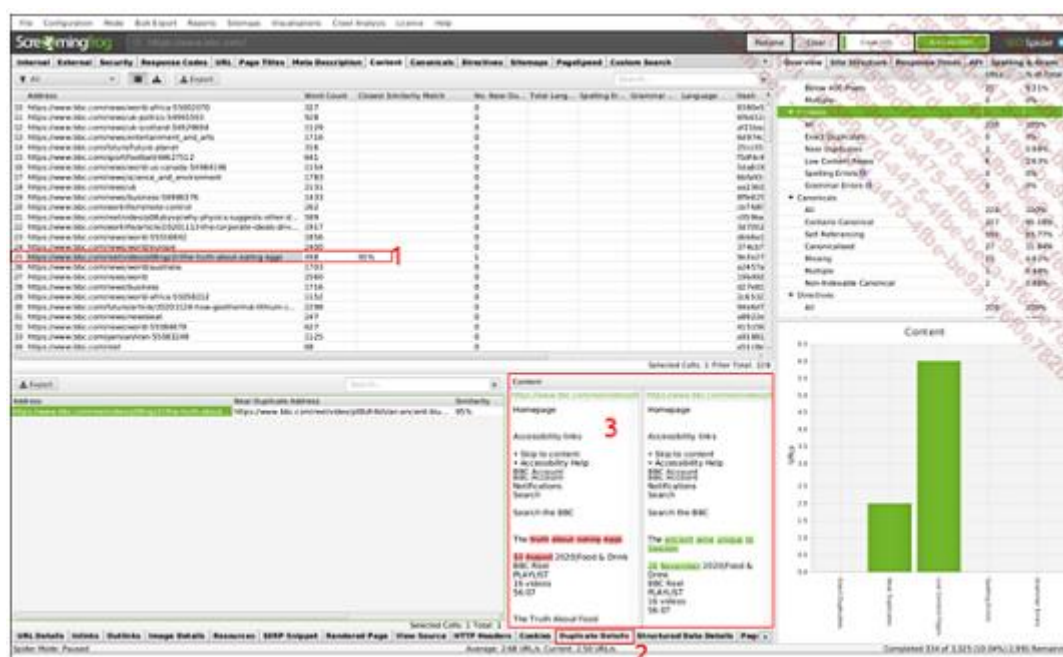
Pour générer les analyses de contenu dupliqué proche, allez dans le menu **Analysis - Configure** et cliquez sur **Content**.



D'autres éléments d'analyse sont également présents dans ce menu, mais ne nous intéressent pas ici, bien que vous pourriez les cocher d'une seule traite pour avoir toutes les informations nécessaires pour des analyses subsidiaires (Hreflang, sitemaps, etc.) que nous verrons plus loin.

Enfin, cliquez sur le bouton **OK**.

Screaming Frog va à présent réaliser les analyses nécessaires. Rendez-vous dans l'onglet **Content** pour les consulter. L'image ci-après vous donne le détail des analyses :



Le point 1 montre une URL qui possède 95 % de similarité avec une autre (le nombre de pages concernées figure dans la colonne intitulée **No. Near Duplicates**). Cliquez sur la ligne à analyser.

Vous pouvez découvrir toutes les URLs similaires en cliquant à présent sur l'onglet du bas intitulé **Duplicate Details** (point 2). L'URL de la page dont le contenu est similaire figure dans les détails de cet onglet. Si plusieurs adresses ont été considérées comme proches en termes de contenu, Screaming Frog affichera alors une ligne par URL.

Le point 3 vous permet de comparer les différences entre les deux pages et de voir en quoi leur contenu est très proche. Dans notre exemple, à peine quelques mots diffèrent (en rouge et vert).

Comme cela est mentionné dans la section Analyser le contenu dupliqué du chapitre Préparation de l'audit sur la configuration de

Screaming Frog, il est possible d'exclure des données du HTML en modifiant certaines zones des pages telles que les menus de navigation. Ce qui est intéressant ici, c'est que ces modifications peuvent se faire à la volée. Voyons cela ci-après.

Changer les zones d'analyse de pages

Imaginez que vous visualisiez les quelques différences entre plusieurs pages et qu'une importante partie de la page ressort (cf. point 3). Dans le menu **Configuration - Content - Area** vous pouvez ajouter la classe CSS correspondant à la partie de la page à faire disparaître de l'analyse. Relancez l'analyse du contenu (menu **Crawl Analysis - Start**). Et voilà : ces parties ont disparu et le rendu HTML est maintenant plus clair.

Export des données

Vous pouvez exporter deux types de rapports :

- La liste des pages dupliquées (bouton **Export** en haut à gauche)
- Le rapport détaillé pour une page en particulier (bouton **Export** en bas à gauche)

Après avoir trouvé ces pages, il sera nécessaire de savoir quelles optimisations leur apporter. Voici quelques pistes :

- Ajouter des balises *canonical* sur les pages de moindre importance, voire créer des redirections 301.
- Faire disparaître les pages inutiles (erreur 404).
- Regrouper du contenu.

9. Analyser les attributs canonical

Cette partie vient compléter l'analyse précédente sur le contenu dupliqué. Le rôle de la balise *canonical* étant précisément d'éviter ce type de problème. Par défaut, on parle de « balise » *canonical*, mais il s'agit en fait d'un attribut de la balise `<link>`.

Rappel sur le fonctionnement de la balise canonical

Son but est d'**éviter le contenu dupliqué interne** (entre des pages d'un même site), mais aussi **externe** (entre différents sites). Si un moteur de recherche arrive sur un site et y trouve une ou plusieurs URLs ayant le même contenu, laquelle doit-il choisir de mettre en avant ? En l'absence de toute indication, il va utiliser plusieurs critères pour déterminer quelle est la page originale (dite page canonique). Il pourra se baser sur la date de première découverte de la page et l'architecture des pages entre elles. Le problème, c'est qu'on ne sait pas s'il va faire la distinction entre la page canonique et ses copies. C'est ici qu'entre en jeu la balise *canonical*.

Située dans l'en-tête HTML d'une page web, elle indique aux moteurs de recherche si la page visitée est celle à prendre en compte ou s'il s'agit de contenu dupliqué. Pour évaluer quels types de pages seront considérées comme canoniques ou comme copies (pages dites canonisées), il vous faudra vous baser sur le contenu lui-même et réfléchir logiquement. Selon les sites et leurs contenus, plusieurs situations sont envisageables.

Code HTML pour une page web :

```
<link rel="canonical" href="https://mon-site.com/categorie/page/" />
```

Les PDF sont également concernés par les balises *canonical*. N'étant pas des pages web, on utilise un en-tête HTTP (qui se configure via le serveur) pour y inscrire le contenu de la balise.

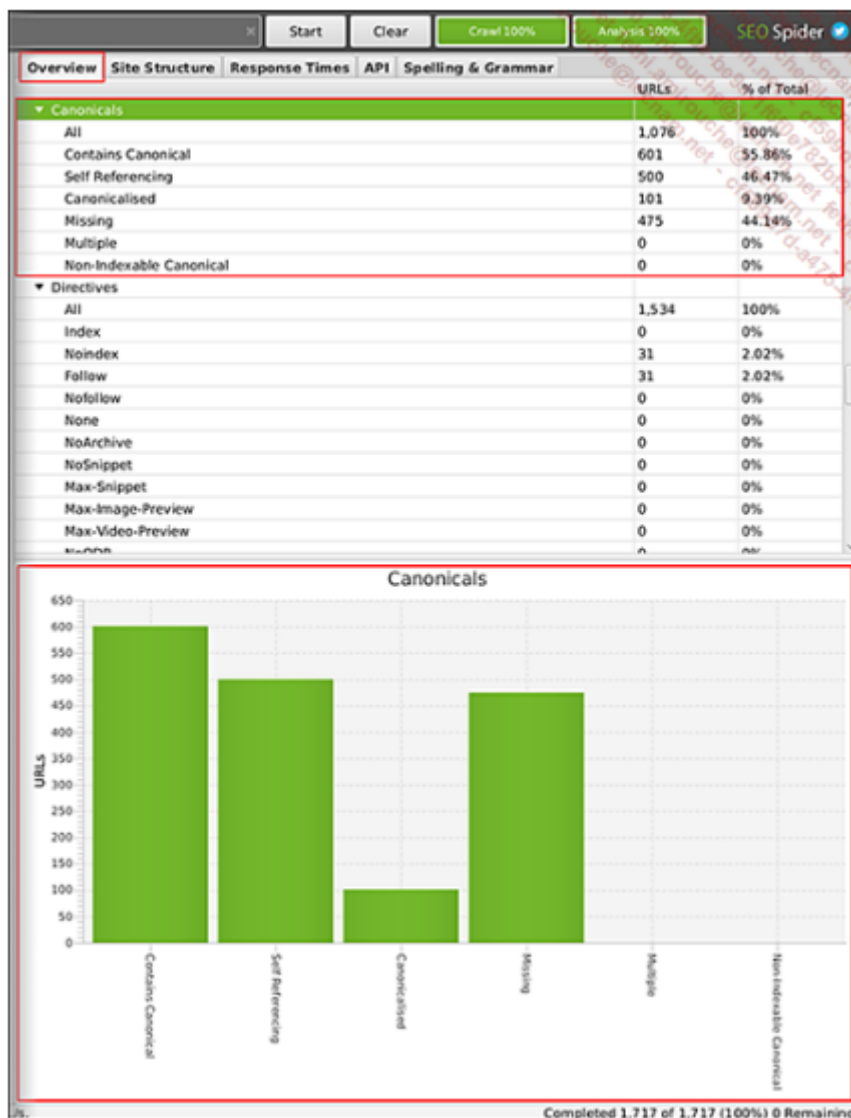
Code pour l'en-tête HTTP d'un fichier PDF (ou autre tel que documents Word, etc.)

```
Link : <link http://mon-site.com/nom-du-fichier.pdf > ;  
rel="canonical"
```

Réaliser une analyse des balises canonical avec Screaming Frog

Le logiciel offre une section dédiée aux balises *canonical* et à leurs problèmes généraux.

Vue globale des balises canonical



Le panneau **Overview** nous propose les éléments suivants :

- **All** : contient l'ensemble des éléments concernés par des balises *canonical*, qu'il s'agisse de pages web ou de fichiers PDF par exemple.
- **Contains Canonical** : donne la quantité d'URL présentant des balises *canonical*/renseignées (aussi bien autoréférentes que des pages canonicisées)
- **Self Referencing** : nombre d'éléments ayant des balises *canonical*/autoréférentes
- **Canonicalised** : nombre d'éléments dont le contenu de la balise *canonical* pointe vers une page canonique. Ce sont les pages canonicisées.

- **Missing** : nombre de pages ou de fichiers sans balise *canonical*.
- **Multiple** : nombre de pages ou de fichiers possédant plusieurs attributs *canonical*.
- **Non-Indexable Canonical** : pages possédant une balise *canonical* (autoréférente ou canonisée) qui ne sont pas indexables (bloquées par le robots.txt, directive nodindex, redirections 301/302/MetaRefresh, pages en erreur 4xx/5xx). N'entrent pas dans cette catégorie les pages canonisées elles-mêmes, bien qu'elles ne soient pas indexables par nature.

Dirigeons-nous maintenant vers les analyses à l'aide des questions suivantes, sur la base de ces informations et en ayant recours à d'autres outils de Screaming Frog.

Y a-t-il des éléments qui devraient avoir une balise canonical mais qui n'en ont pas ?

Il peut s'agir aussi bien de pages web que d'autres types de fichiers tels que des fichiers PDF. Vous trouverez l'information dans la liste des éléments **Missing**.

Les pages canonisées doivent-elles bien être canonisées ?

En principe, une page canonisée renvoie vers une page canonique. Et encore en principe, cette page peut être la page originale d'un contenu. Mais il existe d'autres situations qu'il est primordial de décoder :

- Un renvoi de pages HTTP vers leur équivalent HTTPS. Référez-vous à la section Analyse du crawl et de l'indexation - Statut des éléments crawlés de ce chapitre sur l'analyse des pages indexables pour traiter ce cas.
- Des pages filles qui renvoient vers une page mère. Dans un site e-commerce, ce peut être le cas de pages de description de produits (PDP) détaillées qui renvoient vers une PDP plus générique (ou de PLP). Il s'agit ici de canonisation des *Near Duplicates*. Cette situation est envisageable, mais doit être bien identifiée afin de s'interroger sur sa pertinence, car elle peut ne pas être la plus appropriée (les voies de Google sont parfois impénétrables).

- Des balises *canonical*/peuvent être utilisées à la place des redirections 301. Si vous vous posez la question, retenez qu'une redirection 301 est toujours préférable à une balise *canonical*.

Dans ce cas, analysez les données fournies par l'élément *Canonicalised*.

Vérifier le contenu des balises canoniques

La mise en œuvre de ces balises est le plus souvent automatisée par des plug-ins ou des fonctionnalités propres aux CMS. Pourtant, vous n'êtes pas à l'abri d'un mauvais paramétrage des URLs canoniques, surtout sur les sites e-commerce. Dans Excel, comparez le contenu de la balise canonique avec celui de l'URL de la page, et voyez si des différences apparaissent (manques de parties d'URL par exemple).

Vérifiez notamment que le contenu des balises est bien écrit, de manière relative (ex. : /lunettes-de-soleil.html) comme de manière absolue (https://nom-du-site.com/lunettes-de-soleil.html). Une erreur d'écriture serait d'oublier une partie de l'URL (ex. : nom-du-site.com/lunettes-de-soleil.html).

Les pages d'accueil sont-elles proprement canonisées ?

Cela peut sembler tomber sous le sens, mais sait-on jamais ? Vérifiez que les pages d'accueil possèdent bien **une balise *canonical*/autoréférente**. Les cas d'erreurs sont que la balise pointe vers une autre page d'accueil du site, ou vers une autre page du site sans aucun rapport.

Repérer les pages non indexables

Grâce à l'élément **Non-Indexable Canonical**, vérifiez quelles sont les pages qui disposent d'une balise *canonical*, mais qui ne peuvent pas être indexées. Imaginons une page B qui possède une balise *canonical*/vers une page A, plus importante. Si vous bloquez la page B (intentionnellement ou pas) avec le fichier robots.txt par exemple, Google ne pourra pas savoir que la page B existe et ne pourra pas transférer son PageRank vers la page A. Dommage.

La fonctionnalité présente dans Screaming Frog vous donne des informations pour chaque page de manière individuelle. Mais il peut être fastidieux de vérifier les informations de manière individuelle.

Le mieux est donc d'exporter les données complètes avec les détails des raisons de non-indexabilité grâce au rapport dédié : menu **Reports - Canonical - Non-Indexable Canonicals**.

Corrigez ensuite les pages en fonction.

Les balises canoniques multiples

S'il existe plusieurs balises *canonical* sur une même page, elles seront toutes ignorées par Google ! Rendez-vous dans l'élément **Multiple Canonicals** pour voir quelles sont les URLs concernées.

Le panneau central du logiciel vous permettra de voir quelles sont les multiples références.

Repérer les erreurs courantes de redirection

Dans le principe, l'utilisation de balises canoniques est simple. Parmi les erreurs courantes de redirection, vous pouvez trouver :

- Une page A renvoie vers une page B, qui renvoie elle-même vers la page A.
- Une page A utilise à la fois une balise *canonical* vers une page B et fait une redirection 301 vers cette même page (il faut choisir : canonical ou redirection).
- Les chaînes de pages canonisées : une page A qui renvoie vers une page B qui renvoie vers une page C, etc.

Voyons d'ailleurs comment découvrir les chaînes de redirection grâce à Screaming Frog.

Découvrir les chaînes de redirection dans Screaming Frog :

Rendez-vous au menu **Reports - Canonicals - Canonical Chains**. Exportez les données.

Le tableau vous informe des différents types de chaînes détectées.

Google Search Console à la rescousse

En cas de doute sur les fichiers que Google considère comme canoniques, utilisez l'outil d'inspection d'URL disponible dans Google Search

Console : <https://support.google.com/webmasters/answer/9012289?hl=fr#google-selected-canonical>

Vérifier qu'il n'y a pas de balises canonical dans le corps des pages web

Les balises *canonical* doivent se trouver dans la section `<head>` d'une page web. Elles peuvent aussi se trouver ailleurs, par erreur, notamment dans la section `<body>`. Pour les détecter, mettez en place une recherche personnalisée avant de lancer votre crawl avec Screaming Frog.

Vérifier que les pages canoniques sont bien sur le site

S'il est possible de croiser les URLs pour des sous-domaines différents d'un même site (une page renvoie vers une autre page située sur un sous-domaine du site web), assurez-vous que les liens ne pointent pas vers l'URL d'un template par exemple.

Pour ce faire, dans Excel avec le fichier `nomdedomaine.tld-PAGES.xlsx`, filtrez la colonne **Canonical Link Element 1** en cherchant les URLs qui ne contiennent pas le nom de domaine du site étudié.

Pour de plus amples renseignements sur ces balises, rendez-vous sur la documentation officielle de Google

: <https://developers.google.com/search/docs/advanced/crawling/consolidate-duplicate-urls?hl=fr>

10. Analyser la profondeur des pages

Définition

La profondeur des pages correspond au nombre de clics nécessaires pour atteindre une page donnée depuis la page d'accueil du site. Une profondeur de page trop élevée présente deux problèmes.

- **D'un point de vue de l'UX** (eXpérience Utilisateur, qui définit la qualité du vécu de l'utilisateur) tout d'abord, les utilisateurs devront réaliser davantage de clics pour atteindre le contenu de leur choix, ce qui peut engendrer frustration et perte des visiteurs au passage (surtout si la navigation sur mobile présente de pauvres performances). Google prend d'ailleurs en compte ces types d'éléments comme critères de sélection.
- **Du point de vue du SEO technique**, les pages trop profondes ont plus de risque de ne pas être crawlées. En effet, les moteurs de recherche essaient de préserver leur budget crawl et ces pages bénéficient d'un PageRank plus faible que des pages d'un niveau supérieur. Ceci arrive le plus souvent pour des sites qui présentent **des freins techniques** tels que l'utilisation d'AJAX.

À retenir : les pages trop profondes ont moins de chance d'être découvertes par les moteurs de recherche.

Intérêt de cette analyse

Le but de cette analyse est d'obtenir un ordre de grandeur des pages situées trop profondément dans l'architecture du site. Vous pourrez ensuite vous baser sur la liste des URLs fournies par Screaming Frog pour retravailler leur positionnement dans le site. Selon les sites, il peut s'agir d'une opération assez conséquente où il faudra remanier les pages entre elles en regroupant par exemple du contenu et en utilisant des techniques de hiérarchisation telles que des grappes de contenu (ou de cocon sémantique). Un inventaire du contenu pourrait être nécessaire mais ceci n'est pas le propos de cet ouvrage.

Sur des sites de taille plus modeste, il ne s'agit pas de faire remonter les pages de manière artificielle pour qu'elles se retrouvent à un niveau de profondeur plus faible, mais bien de les mettre en avant à travers des liens vers la homepage ou vers d'autres rubriques.

Causes probables d'une profondeur de pages trop élevée

Plusieurs scénarios sont possibles :

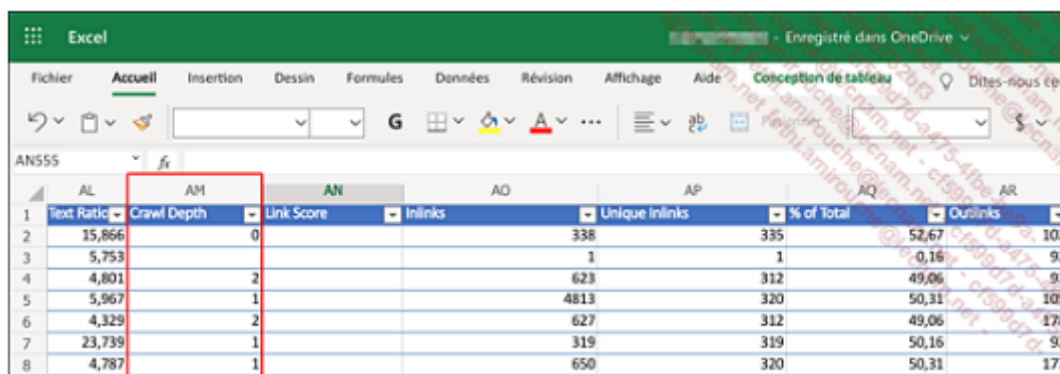
- **Mauvaise hiérarchisation de l'information** : la structure des pages a été mal pensée, sans stratégie préalable ou planification.
- **Paramètres d'URL trop importants** : des paramètres sont ajoutés aux URLs, pour analyser par exemple les mouvements des utilisateurs sur le site (page source, etc.), créant une infinité d'URLs possibles.
- **Paramétrage des filtres sur des sites e-commerce** : il s'agit de la même situation que le point précédent, mais elle mérite un intérêt particulier car elle est relativement courante. Le filtrage des catégories de produits (par taille, couleur, public concerné, etc.) peut ainsi se retrouver dans les URLs, ce qui entraîne ici aussi la création d'une multitude d'URLs. Pour peu que des éléments de navigation, voire de tracking du niveau de scroll dans la page se retrouvent dans les URLs, vous imaginez la catastrophe. Ce type de site engendre souvent un fichier de crawl de taille importante, et ne permet pas d'atteindre un crawl de 100 %, les possibilités d'URL étant quasi illimitées. Il faut alors retraiter les données du site pour trouver les URLs réellement exploitables et intéressantes. À ce titre, l'utilisation de AWK peut vous être d'un grand secours, Excel est parfois limité dans le traitement des fichiers d'exports de plusieurs dizaines de Mo.
- **Des URLs mal formées** : c'est également une spécialité des sites e-commerce, pour lesquels on peut retrouver des pages accessibles via plusieurs URLs en raison du paramétrage des filtres.

Un exemple : /lunettes-de-vue/homme/gucci
vs/homme/lunettes-de-vue/gucci/vs/gucci/lunettes-de-vue/homme/vs/lunettes-de-vue/gucci/.

Dans ce cas, on obtient quatre URLs possibles pour afficher une même page produit (PLP de la marque Gucci). C'est d'ailleurs un cas typique de DUST (*Duplicate URL, Same Content*). Le DUST désigne une situation où une même page est accessible à travers des URLs différentes, engendrant du contenu dupliqué.

- **Une pagination mal conçue** qui limite la vue sur les trois prochaines pages du site, alors que vous avez une soixantaine d'éléments à montrer. Le nombre de clics pour atteindre les dernières pages est donc très élevé.
- **Des liens présents sur toutes les pages**, comme sur une page contact (« Voir contact précédent » et « Voir contact suivant »), créant ici aussi une infinité de possibilités.

Analyse globale de la situation

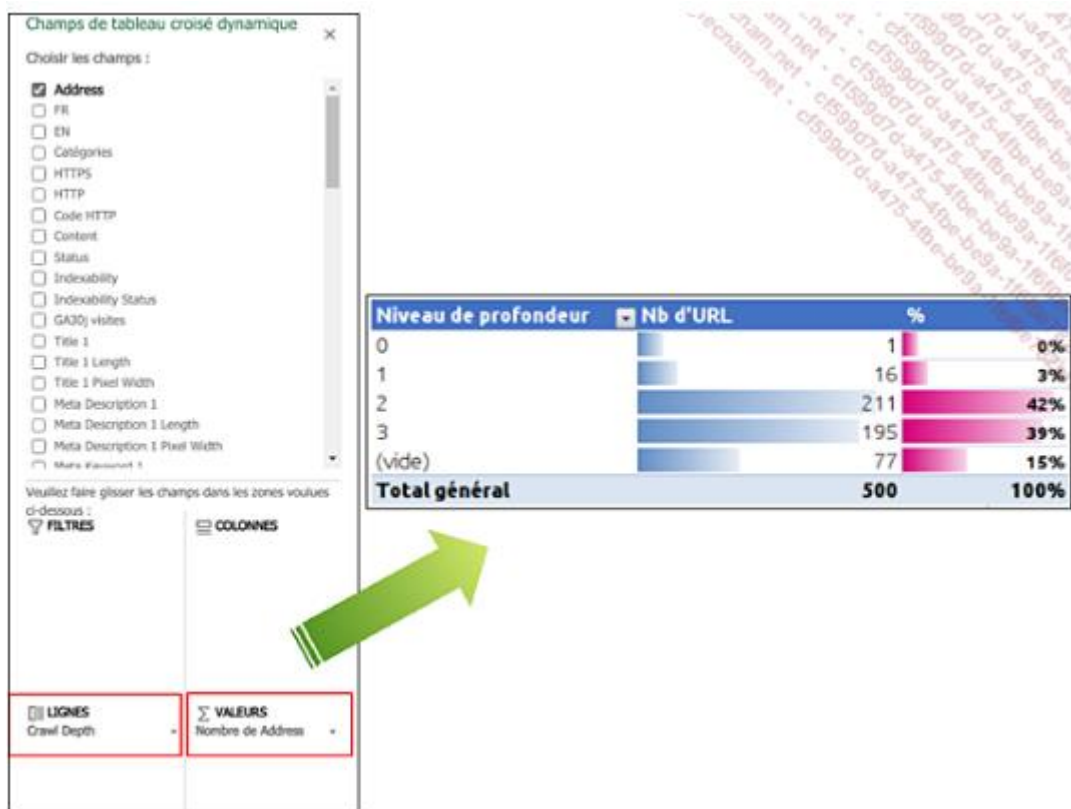


	AL	AM	AN	AO	AP	AQ	AR
	Text Ratio	Crawl Depth	Link Score	Inlinks	Unique Inlinks	% of Total	Outlinks
2	15,866	0			338	335	52,67
3	5,753				1	1	0,16
4	4,801	2			623	312	49,06
5	5,967	1			4813	320	50,31
6	4,329	2			627	312	49,06
7	23,739	1			319	319	50,16
8	4,787	1			650	320	50,31

Nous allons utiliser Excel et le fichier nomdedomaine.tld-PAGES.xlsx et utiliser les tableaux croisés dynamiques. Le fichier d'export des URLs de Screaming Frog présente une colonne intitulée **Crawl Depth** qui contient **la profondeur**, en nombre de clics, de chaque URL découverte par le crawler.

Dans cet exemple, la colonne apparaît sous le nom AM. Initialement, cette référence est moins avancée, mais la création de nouvelles colonnes (HTTPS, HTTP, etc.) modifie la classification de départ du fichier.

Paramétrage du tableau croisé dynamique



Utilisez les paramètres suivants :

- **LIGNES** : placez le champ **Crawl Depth**.
- **VALEURS** : placez le champ **Address** en mode Nombre.

Le tableau présente le nombre d'URL par niveau de profondeur. Créez une colonne avec les pourcentages et, comme pour chaque tableau croisé dynamique, optez pour une mise en valeur des données à l'aide de la mise en forme conditionnelle. Vous pouvez bien sûr créer un graphique, à barres de préférence, pour mettre en exergue cette analyse dans votre rapport d'audit.

Analyse approfondie

Dans notre analyse, nous voyons apparaître différents éléments. Nous allons les expliquer pour ce cas précis, ce qui vous servira de base pour vos propres analyses.

Le niveau 0 correspond à une redirection 301 de la page d'accueil générique vers une page se trouvant dans le sous-répertoire /en/ du site. Il n'est pas vraiment important, les moteurs de recherche comprennent aisément cette manipulation. Ce sont les niveaux suivants qui nous intéressent.

Le site possède trois niveaux de profondeur, la plupart des pages se trouvant dans les niveaux 2 et 3. Sur le site analysé, c'est relativement sain. De manière générale, le niveau 3 représente une barrière à ne pas dépasser. Cela dépend bien sûr du type de site. Un site e-commerce mettant en avant plus de sous-catégories de produits pourra avoir des niveaux de profondeur plus élevés, ce qui serait tout à fait acceptable pour des pages produits (PDP) par exemple. Cependant, des optimisations seraient éventuellement à prévoir pour mettre en avant les pages de catégorie produits.

Notre exemple montre également des pages ayant un niveau de profondeur vide. Comment se fait-il que des pages qui n'ont pas pu être crawlées se retrouvent dans cette analyse ? La réponse est simple : ce sont des **pages orphelines**. Elles ont bien été remontées par les données fournies par Google Analytics, les sitemaps ou Google Search Console, mais elles ne sont liées à aucune autre page du site. Il sera évidemment intéressant de vous référer à la section Analyser les sitemaps de ce chapitre relative aux pages orphelines pour les traiter.

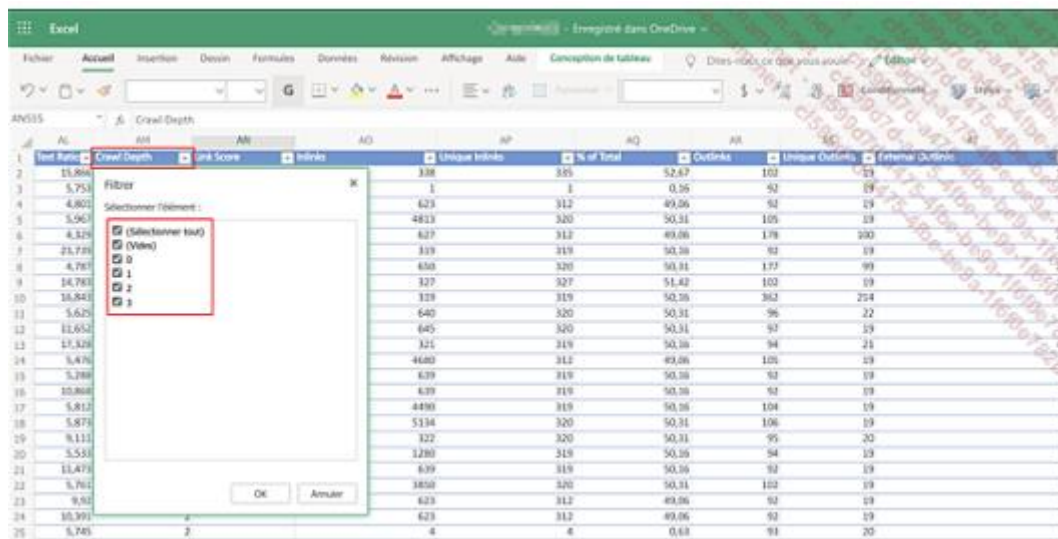
Découvrir les pages par niveau de profondeur

Vous avez plusieurs possibilités tant avec Excel qu'avec Screaming Frog :

- Possibilité 1 : en triant les données directement dans le tableau Excel.
- Possibilité 2 : en filtrant les données grâce au tableau croisé dynamique.
- Possibilité 3 : en filtrant les données dans Screaming Frog.

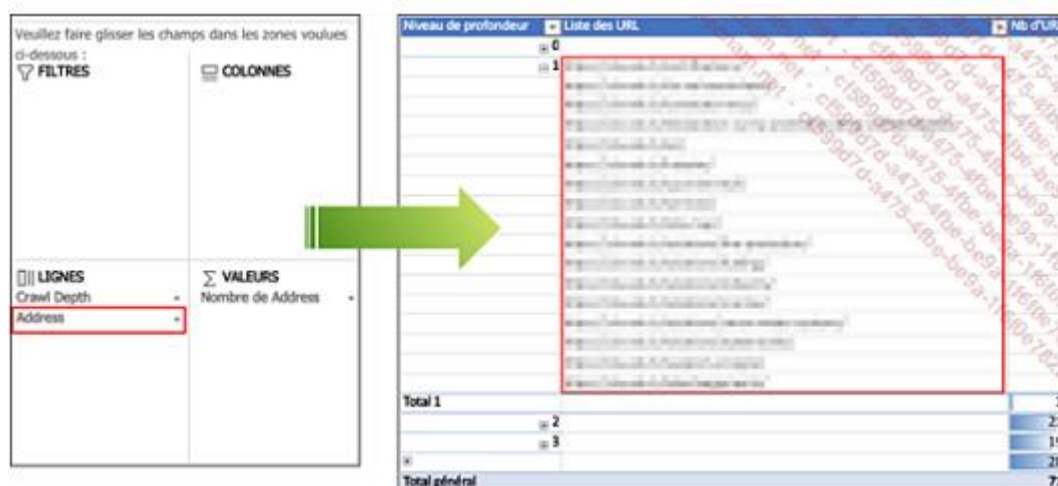
Possibilité 1 : en triant les données directement dans le tableau Excel

C'est une solution relativement simple. Dans la feuille de calcul des URLs, filtrez les données par niveau de profondeur puis récupérez la liste de ces URLs. Cette opération est à faire pour les URLs se trouvant dans les niveaux de profondeur les plus élevés, quoique rien ne vous empêche de vérifier l'ensemble des URLs, à la recherche de causes probables.



Possibilité 2 : en filtrant les données grâce à un tableau croisé dynamique

Il suffit de rajouter le champ **Address** dans la section **LIGNES** du tableau croisé dynamique précédent pour voir apparaître les URLs selon leur niveau de profondeur.



Voici une autre solution, tout aussi rapide, à l'aide d'un tableau croisé dynamique :

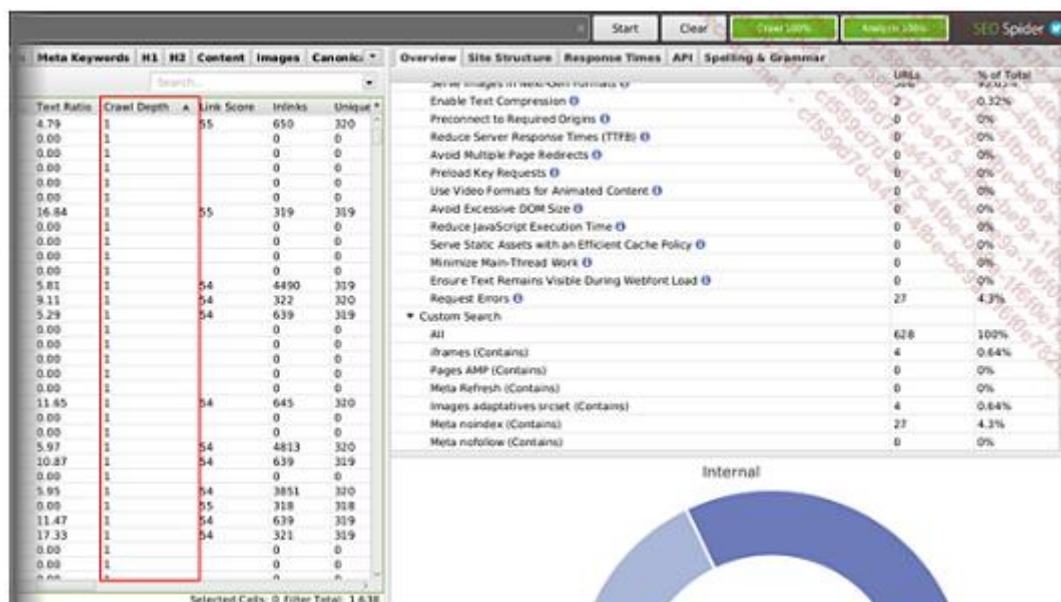
Créez un nouveau tableau croisé dynamique basé sur le tableau situé dans la feuille de calcul URL.

Dans la section **LIGNES**, choisissez le champ **URL**.

Dans la section **FILTRES**, choisissez le champ **Crawl Depth**.



Possibilité 3 : en filtrant les données dans Screaming Frog



Analyse avec pages actives et inactives

1. Introduction

Nous entrons ici dans une des parties les plus enrichissantes de notre analyse. Nous allons **coupler les données du crawl Screaming**

Frog avec des données extraites de Google Analytics. Au début de cette analyse, nous avons configuré notre crawler avec la solution de web analyse de Google, ce qui nous a permis de trouver certaines informations intéressantes, comme par exemple les pages orphelines.

Maintenant, nous allons pouvoir aller plus loin dans nos investigations. Et le cœur du système va être notre tableau d'URLs sous Excel, comme cela a été abordé dans l'introduction du présent ouvrage.

À partir des données de Google Analytics, nous avons déjà des éléments de réponse sur les points suivants :

- **Les pages actives** du site.
- **Les pages inactives du site** qui ne reçoivent pas de trafic.
- Les catégories de **pages les plus visitées**.
- Les principaux indicateurs ou **KPI** sur différents types de pages : taux de rebond, durée moyenne de visite, etc.

Nous pourrons enrichir nos futures analyses, notamment on-page avec ces données (pour rappel le SEO on-page fait référence à tout ce qui est modifiable directement sur le site (balise, contenu, sitemap, maillage interne, etc.)). Les manipulations suivantes nécessitent une certaine maîtrise de **Google Analytics**. Vous ne trouverez pas d'explication sur son fonctionnement ni ses concepts clés. Entrons maintenant dans le vif du sujet.

2. Télécharger les données de Google Analytics

Avant de commencer

Une petite mise en garde concernant **la qualité des données** remontées dans Google Analytics s'impose au préalable. Sur un site important, il y a peu d'inquiétude à avoir, les équipes marketing font souvent appel à des professionnels du web analytics. Leurs comptes sont structurés correctement, et les vues sont également filtrées pour éviter tout trafic parasite.

Sur des comptes d'organisations plus modestes, il arrive que les comptes soient mal structurés ou mal paramétrés. Si un compte existe, il se peut que seule la vue **Toutes les données** soit présente et sa propriété non paramétrée. Attention également au paramétrage des TMS (*Tag Management System*) comme Google Tag Manager, une mauvaise manipulation sur la simple balise de remontée de Google Analytics peut bloquer une partie de la collecte des informations de tracking.

Récupération des données de trafic organique dans Google Analytics

Accédez à votre compte Google Analytics. Si le compte dispose d'une vue principale bien paramétrée, choisissez-la. Sinon, utilisez la vue **Toutes les données**.

Choix du rapport : dans Google Analytics (Universal Analytics), allez dans le rapport suivant : **Comportement - Contenu du site - Toutes les pages**

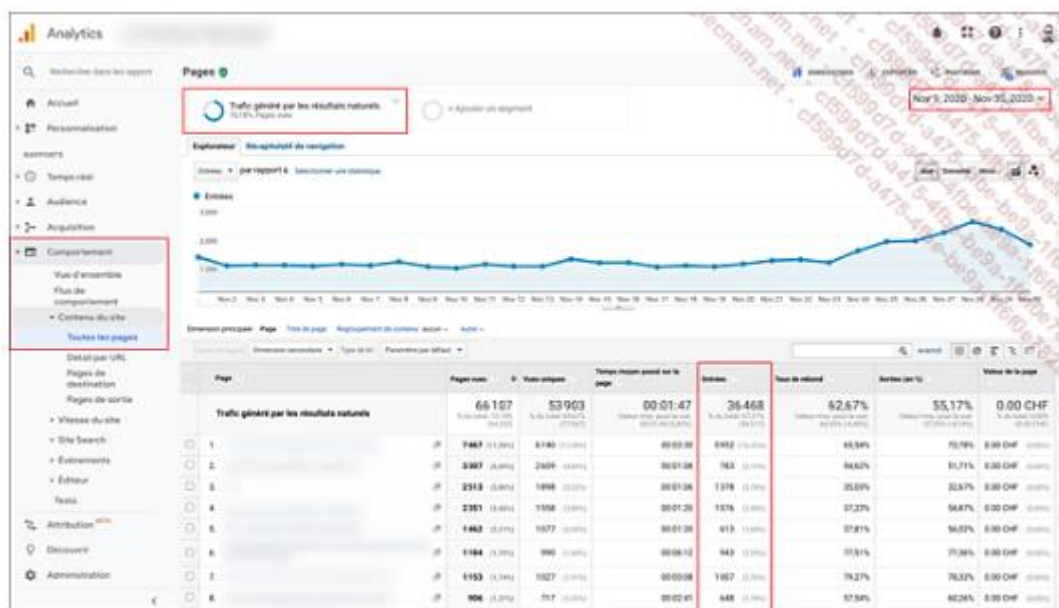
Choix du segment : supprimez le segment **Tous les utilisateurs** et remplacez-le par le segment **Trafic généré par les résultats naturels**.

Périodes d'analyse : nous allons réaliser deux analyses, la première portant sur les 30 jours précédant le crawl, la deuxième portant sur les 90 jours précédant le crawl.

Si vous avez oublié la date du crawl, vous pouvez la retrouver dans Screaming Frog dans le menu **Reports - Crawl Overview**.

Nous allons nous intéresser à la colonne **Entrées** présente dans ce rapport. En choisissant le segment relatif au trafic organique, nous obtenons les entrées organiques, c'est-à-dire les entrées sur le site obtenues par des clics sur des liens présents dans les SERP. Les pages qui apparaissent ici sont donc les pages d'atterrissages organiques.

Ce rapport nous présente également les KPI les plus importants rencontrés de manière générale dans Google Analytics (taux de rebond, temps moyen, sorties, etc.).



Export des données : une des particularités de Google Analytics est de limiter l'export aux données présentées dans la vue courante. Affichez donc l'ensemble des lignes du rapport au risque de vous retrouver avec à peine 10 lignes.

Puis cliquez sur le bouton **Exporter** et choisissez le format Excel (XLSX) pour pouvoir traiter les données plus rapidement ensuite. Vous pouvez notamment exporter ce rapport aux formats CSV et Google Sheets comme bon vous semble.

Préparation des fichiers

Import des données dans Excel

Ouvrez le fichier téléchargé.

Comme tout rapport de Google analytics, il présente plusieurs feuilles de calcul, et c'est la deuxième feuille de calcul, sobrement appelée **Ensemble des données 1**, qui nous intéresse. Il contient les informations présentes dans le rapport Google Analytics.

Ouvrez le fichier Excel nomdedomaine.tld-PAGES.xlsx et importez la feuille de calcul issue de Google Analytics dans une nouvelle feuille de calcul. Renommez cette feuille de calcul **GA30j** (pour Google Analytics 30 jours). Vous pouvez évidemment choisir la nomenclature qui vous va le mieux, mais dans le cadre de ce livre, nous resterons sur cette version par souci de clarté.

Recommencez la même opération sur Google Analytics, mais en vous basant sur les 90 derniers jours précédents le crawl du site. Renommez la feuille de calcul finale **GA90j**.

Formatage du tableau et des URLs

Donnez un peu de couleur à votre **feuille de calcul GA30j** en la mettant en forme (**Mettre sous forme de tableau**) et donnez au tableau le même nom que celui de la feuille de calcul (GA30j). Cette dernière manipulation sera très pratique par la suite. Pour nommer un tableau sous Excel, et comme vous avez fait pour le tableau des URLs du crawl, placez-vous dans une cellule du tableau, puis allez dans l'onglet **Conception de tableau** (ou **Outils de tableaux - Création** selon votre version d'Excel) et remplissez le champ **Nom du tableau** situé en haut à gauche dans le ruban.

Au préalable, il est nécessaire de formater les URLs de la première colonne de Google Analytics, ce dernier exportant en fait les **URI** (*Uniform Resource Identifier*, c'est-à-dire l'adresse d'une page web sans la partie <https://www.nom-du-site.com>). Pour cela, nous allons ajouter la partie manquante.

Créez une nouvelle colonne à droite, en deuxième position, et utilisez la formule CONCATENER qui va vous permettre, comme son nom l'indique, de regrouper plusieurs éléments.

Sur Excel en français, la formule à appliquer sur la première ligne de données (pas l'en-tête) est la suivante : **=CONCAT("https://nom-du-site.com" ; A2)**. Du fait de la mise en forme appliquée précédemment, elle va se répercuter sur toutes les cellules qui suivent.

Faites un copier-coller des valeurs de cette colonne dans la colonne **Page** initiale. Vous avez à présent une colonne contenant les valeurs des URLs (pas les URI de Google Analytics ni les cellules avec la formule CONCATENER).

Répétez la même opération pour les données sur les 90 jours précédant le crawl dans la **feuille de calcul GA90j** et nommez le tableau GA90j.

	A	B	C	D	E	F	G
	Page	Pages vues	Vues uniques	Temps moyen	Entrées	Taux de rebond	Sorties (en %)
1	https://www.nom-du-site.com/	39	33	15,06	32	15,63%	20,51%
2	https://www.nom-du-site.com/	15	9	16,69	1	0,00%	13,33%
3	https://www.nom-du-site.com/	14	14	24,00	10	100,00%	92,86%
4	https://www.nom-du-site.com/	11	10	39,50	0	0,00%	81,82%
5	https://www.nom-du-site.com/	8	7	30,43	0	0,00%	12,50%
6	https://www.nom-du-site.com/	7	6	21,17	1	0,00%	14,29%
7	https://www.nom-du-site.com/	4	3	14,00	0	0,00%	25,00%
8	https://www.nom-du-site.com/	4	4	230,00	2	100,00%	75,00%
9	https://www.nom-du-site.com/	2	2	370,00	1	100,00%	50,00%
10	https://www.nom-du-site.com/	2	1	38,00	0	0,00%	50,00%
11	https://www.nom-du-site.com/	1	1	32,00	0	0,00%	0,00%
12	https://www.nom-du-site.com/	1	1	10,00	0	0,00%	0,00%
13	https://www.nom-du-site.com/	1	1	8,00	0	0,00%	0,00%
14	https://www.nom-du-site.com/	1	1	0,00	0	0,00%	100,00%
15	https://www.nom-du-site.com/	1	1	9,00	0	0,00%	0,00%
16	https://www.nom-du-site.com/	1	1	4,00	0	0,00%	0,00%
17	https://www.nom-du-site.com/	1	1	20,00	0	0,00%	0,00%
18	https://www.nom-du-site.com/	1	1	0,00	0	0,00%	100,00%
19	https://www.nom-du-site.com/	1	1	12,00	1	0,00%	0,00%
20	https://www.nom-du-site.com/	1	1	249,00	1	0,00%	0,00%
21	https://www.nom-du-site.com/	1	1	94,00	0	0,00%	0,00%
22	https://www.nom-du-site.com/						

3. Pages actives vs Pages inactives vs Pages fantômes

Définition

Une page active est une page qui reçoit au moins une visite au cours des 30 jours précédant le crawl. Il est possible d'être même un peu plus sévère sur un site à fort trafic et d'estimer cette mesure à deux visites sur les 30 derniers jours.

Dès lors, une page est dite **inactive** lorsqu'elle ne reçoit aucune visite au cours de cette même période de 30 jours.

Mais le plus important ce sont **les pages fantômes** qui, pour leur part, n'ont reçu aucun trafic durant les 90 jours qui précèdent le crawl. **Elles sont un bon indice d'un problème d'indexation.**

Analyse des pages

Pour réaliser cette étude, vous avez besoin d'avoir importé et mis en forme les données sur les deux périodes mentionnées précédemment (feuilles de calcul GA30j et GA90j).

Dans la **feuille de calcul URL** qui est la première feuille de calcul de votre fichier Excel et qui contient les données du crawl Screaming Frog, créez deux colonnes et nommez-les respectivement **GA30j** et **GA90j**.

À l'aide de la formule RECHERCHEV, nous allons à présent croiser les données issues de la colonne **Entrées** des rapports Google Analytics avec les données du crawl.

Dans la première cellule de la colonne GA30j, saisissez la formule suivante :

=RECHERCHEV(A2;GA30j;5;0)

Le calcul se répercute sur toutes les lignes qui suivent ; le résultat est le nombre d'entrées sur les pages qui ont été visitées. Les pages n'ayant pas reçu de visites sont affublées de la mention #N/A (signifiant *Not Available*, c'est-à-dire *Non Disponible*).

Répétez la même opération sur la colonne d'à côté portant sur les 90 jours précédents en utilisant la formule suivante :

=RECHERCHEV(A2;GA90j;5;0)

URL	GA30j	GA90j	Title	Title Length	Title Pixel	Meta Description 1	Meta Description 2	Meta Description 3
	32	104			46	424	151	964
	#N/A	#N/A			29	270	0	0
	#N/A	0			22	193	20	140
	1	2			72	692	149	905
	10	13			19	180	86	563
	#N/A	#N/A			60	687	92	598
	0	6			58	563	129	839
	0	10			151	1333	8	45
	#N/A	#N/A			18	161	0	0
	#N/A	#N/A			55	512	155	803
	#N/A	0			67	618	153	981
	#N/A	#N/A			22	195	90	585
	1	5			20	190	121	802
	#N/A	#N/A			58	537	79	505
	#N/A	#N/A			61	569	153	1017
	#N/A	#N/A			64	574	155	994
	#N/A	#N/A			79	704	132	842
	0	0			56	507	93	675
	#N/A	0			83	716	132	827
	2	3			40	362	153	1006
	#N/A	0			72	641	134	842
	#N/A	0			26	266	120	795
	#N/A	#N/A			17	144	149	950
	1	3			60	533	121	756
	#N/A	0			17	151	93	675
	#N/A	#N/A			18	161	0	0
	#N/A	#N/A			38	429	79	492
	#N/A	#N/A			19	170	86	558
	#N/A	0			24	203	57	370
	#N/A	#N/A			18	149	144	894
	0	0			56	497	0	0
	#N/A	#N/A			16	143	143	875
	#N/A	#N/A			27	227	148	930
	0	2			66	623	151	990

V pour Vertical : lumière sur la formule RECHERCHEV

Avec cette formule, on demande à Excel d'aller chercher le contenu de la cellule A2, qui est la cellule contenant la première URL crawlée par Screaming Frog, dans les tableaux de données de Google. Excel va ensuite afficher le résultat de la cellule située à la cinquième colonne correspondant à la même ligne que celle de l'URL trouvée dans ces tableaux. Le V de RECHERCHEV signifie Vertical,

Excel fait une recherche du contenu de la première cellule en descendant dans le tableau, donc de manière verticale.

Catégorisation des pages selon les visites

Nous allons catégoriser les pages aussi selon qu'elles sont des pages actives, inactives ou bien fantômes.

À la suite des colonnes GA30j et GA90j, créez deux nouvelles colonnes intitulées respectivement :

- Statut Pages 30 j
- Pages Fantômes

Pour remplir les cellules de la colonne **Statut Pages 30 j**, nous allons filtrer la colonne GA30j deux fois :

Un premier filtrage pour ne garder que les valeurs supérieures à 0 et en excluant les valeurs #N/A. Remplissez alors les cellules correspondantes de la colonne **Statut Pages 30 j** avec la mention **Pages actives**.

Un second filtrage en ne gardant que les valeurs égales à 0 ou à #N/A. Remplissez cette fois les cellules de la colonne **Statut Pages 30 j** avec la mention **Pages inactives**.

Pour remplir la colonne **Pages Fantômes**, filtrez la colonne GA90j sur les valeurs égales à 0 ou à #N/A. Saisissez *Pages fantômes* dans les cellules correspondantes.

En résumé, vous obtenez :

- d'une part une colonne intitulée **Statut Pages 30 j** qui comprend le statut des pages sur les 30 derniers jours précédant le crawl (pages actives et pages inactives),
- une autre colonne intitulée **Pages Fantômes** qui renseigne sur le statut des pages sur la période de 90 jours précédant le crawl.

	L	M	N	O	P	Q
1	Indexability Status	GA30j	GA90j	Statut Pages 30j	Pages Fantômes	Titre 1
2			38	100	Pages actives	
3		#N/A	#N/A	Pages inactives	Pages fantômes	
4			0	0	Pages inactives	Pages fantômes
5			1	2	Pages actives	
6			9	11	Pages actives	
7		#N/A	#N/A	Pages inactives	Pages fantômes	
8			2	7	Pages actives	
9	Canonicalised		0	12	Pages inactives	
10		#N/A	#N/A	Pages inactives	Pages fantômes	
11		#N/A	#N/A	Pages inactives	Pages fantômes	
12			0	0	Pages inactives	Pages fantômes
13		#N/A	0	0	Pages inactives	Pages fantômes
14			1	1	Pages actives	
15		#N/A	#N/A	Pages inactives	Pages fantômes	
16		#N/A	#N/A	Pages inactives	Pages fantômes	
17		#N/A	#N/A	Pages inactives	Pages fantômes	
18		#N/A	#N/A	Pages inactives	Pages fantômes	
19			0	0	Pages inactives	Pages fantômes
20		#N/A	0	0	Pages inactives	Pages fantômes
21			2	3	Pages actives	
22			0	0	Pages inactives	Pages fantômes
23		#N/A	0	0	Pages inactives	Pages fantômes

Quid des cellules avec un 0 et non #N/A

Ce sont simplement des cellules qui ont reçu du trafic interne ou externe (*backlinks*), des visites en provenance d'autres pages, mais qui n'ont pas servi de porte d'entrée au site depuis un moteur de recherche.

Les cellules ayant #N/A sont inconnues de Google Analytics ! Elles n'ont reçu ni visites depuis un moteur de recherche, ni visites depuis une autre page du site ou externe.

Analyse des données avec Excel

Nous allons de nouveau réaliser un tableau croisé dynamique qui va nous permettre de vérifier la part des pages actives et inactives, ainsi que celles des pages fantômes. Nous pourrions voir quelles sont les pages qui méritent d'être mises en avant.

Paramétrez Excel pour créer un tableau dynamique :

- **LIGNES** : Statut Pages 30 j
- **VALEURS** : Address (en mode Nombre) et Pages Fantômes (en mode Nombre)
- **FILTRES** (optionnel) : Indexability

Mettez en forme le tableau généré et renommez l'en-tête de la colonne Nombre de Address en **Nombre d'URL**.

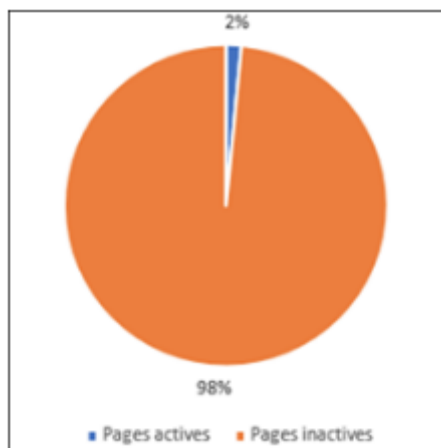
Statut Pages 30j	Nombre d'URL	Nombre de Pages Fantômes
Pages actives	11	
Pages inactives	706	692
Total général	717	692

Ce tableau indique **une vue générale de la situation**. En ajoutant un filtre avec la colonne **Indexability**, il est possible de ne sélectionner que les pages indexables et ainsi d'avoir une vision assez proche de celle d'un moteur de recherche (qui n'affiche que les pages indexables).

Aussi il est tout à fait possible de gérer ces données en deux parties distinctes :

- Pages actives vs pages inactives
- Pages fantômes vs total des pages

Vous pourrez notamment créer des graphiques pour chacune des parties comme celui-ci :



Préconisations

Au vu de ces résultats, le travail d'optimisation peut se focaliser sur les pages qui génèrent le plus d'activité, c'est-à-dire les pages actives.

Pour calculer le nombre de visites organiques sur ces pages, il suffit de faire la somme de la colonne GA30j dans la feuille de calcul URL, ou de faire un tableau croisé dynamique... Cette dernière idée est

quelque peu « overkill » comme on dit en agence web, mais je ne résiste pas à vous la partager.

Indexability	Indexable	
Statut Pages 30j	Nombre de GA30j	Somme de GA30j
Pages actives	11	58
Pages inactives	489	#N/A
Total général	#N/A	#N/A

Dans notre cas, 11 pages ont généré 58 visites sur les 30 derniers jours (le site n'est pas très productif). Il suffit d'extraire leurs URLs en filtrant la colonne GA30j sur **Pages actives**.

Il peut également être intéressant de voir pourquoi les autres pages ne génèrent pas de visites.