Gephi et l'analyse de données pour le SEO (1ère partie)

Par Daniel Roch

Domaine :	**Recherche**	**Référencement**
Niveau :	Pour tous	**Avancé**

Gephi est un logiciel dont on entend parler régulièrement pour analyser la structure d'un site web. En réalité, Gephi est un logiciel d'analyse capable de traiter et de visualiser n'importe quel type de données, permettant ainsi d'approfondir notre analyse en référencement naturel. Nous allons donc voir dans cet article en deux parties comment ce logiciel fonctionne, ses qualités (mais également ses défauts) et comment on peut en tirer profit dans le cadre d'une stratégie SEO.

Depuis quelques années, les référenceurs se sont rendu compte que l'outil Gephi permettait de visualiser des informations liées au référencement naturel d'un site, notamment la présence de silos ou encore la popularité des pages, et ainsi pouvoir mieux agir et prendre certaines décisions.

Dans cette première partie, nous allons apprendre à utiliser l'outil et à trier, classer et visualiser les données. Le mois prochain, nous aborderons des cas concrets d'analyses SEO en utilisant Gephi.

Qu'est-ce que Gephi ?

Gephi est un logiciel de visualisation spatiale de données. Il est très souvent utilisé dans les domaines liés à l'algorithmie et aux statistiques. Il s'agit d'un logiciel Open Source gratuit, disponible sur Windows, Mac OS X et Linux. Il fonctionne avec Java (version 7 minimum) et peut nécessiter un ordinateur puissant pour le traitement des données (que ce soit au niveau du processeur ou de la mémoire vive disponible).

Le site officiel pour le télécharger se trouve ici : https://gephi.org/

Fig. 1. Le logo de Gephi

Point intéressant, on peut y ajouter des plugins pour différentes fonctionnalités supplémentaires. Ils sont tous disponible à cette adresse : https://marketplace.gephi.org/

Autre élément à connaître : la version officielle actuelle ne marche pas totalement, et une des fonctionnalités nécessaire pour visualiser les silos ne fonctionne plus. Il est donc conseillé pour suivre ce guide d'installer la version « nightly build » 0.9.2 : https://github.com/gephi/gephi#nightly-builds

Remarques préalables

Avant de vous lancer dans l'utilisation du logiciel Gephi, il est important de savoir plusieurs choses. La première est que Gephi est parfois instable, surtout lors du traitement de données variées et nombreuses. Pensez donc à sauvegarder souvent votre travail !

Autre point problématique : il faut avoir conscience que Gephi ne dispose pas de fonction « Annuler ». Une action menée à son terme est donc irréversible. Nous conseillons ainsi non seulement de sauvegarder souvent, mais aussi de le faire à chaque fois dans des fichiers différents, par exemple gephi-01, gephi-02, gephi-03…

Du fait de son instabilité, il est parfois conseillé d'importer uniquement les données qui seront utilisées, quitte à refaire une importation lorsqu'on voudra analyser un point différent. Dans les exemples que nous donnerons, nous avons justement trié les données en amont pour n'importer que les URL de pages HTML.

Ce qui nous amène au point le plus important de l'utilisation de Gephi dans une optique de référencement naturel : pour faire du bon travail, **il est impératif de savoir ce que l'on veut analyser avant d'utiliser le logiciel**, au risque de perdre du temps à faire une multitude de tests et de paramétrages pour rien.

Fonctionnement de base

Gephi est un logiciel qui visualise des données. On l'utilise notamment dans les analyses statistiques sur des populations, des communautés, sur la mise en avant de leaders d'opinion mais également en SEO. Dans une optique de référencement naturel, Gephi est particulièrement efficace pour l'analyse de la structure d'un site web.

Quel que soit ce que l'on en fait, le fonctionnement du logiciel est le même :

- On récupère des données, que l'on classe en deux fichiers ;
- On importe ces données ;
- On les modifie dans une fenêtre de visualisation ;
- Puis on analyse le rendu final.

1- Collecte de données

La première étape est donc de pouvoir récupérer les données que l'on souhaite analyser. Nous allons devoir, dans un premier temps, crawler le site que l'on veut analyser, et récupérer toutes les informations que l'on voudrait mettre en avant. Pour cela, on peut utiliser différents outils, notamment :

- Xenu Link Sleuth – Gratuit : http://home.snafu.de/tilman/xenulink.html ;
- Screaming Frog Spider SEO – Payant (environ 150 € / an) : https://www.screamingfrog.co.uk/seo-spider/ ;

Dans cet article, nous utiliserons le second car il permet par défaut de récupérer plus de données pour chaque URL, et surtout il peut être connecté lors du scan à Google Analytics et à la Search Console de Google, permettant ainsi nativement pour chaque URL de récupérer des informations liées au Traffic ou à la visibilité.

Ensuite, nous conseillons fortement d'avoir un accès à des données liées aux liens externes vers votre site (backlinks), par exemple avec :

- Majestic SEO : https://fr.majestic.com/ ;
- Open Site Explorer : https://moz.com/researchtools/ose/ ;
- Ahrefs : https://ahrefs.com/dashboard/metrics ;

Et pour terminer, nous conseillons également l'utilisation de l'extension Excel SEO Tools, qui permet notamment de récupérer des données comme le nombre de liens, le nombre de domaines référents, le Trust Flow ou le Citation Flow d'une URL directement dans l'interface d'Excel.

Site officiel de SEO Tools : http://seotoolsforexcel.com/

Nous commençons donc par connecter Screaming Frog à Google Analytics et à la Search Console, puis nous lançons le crawl du site qui nous intéresse.

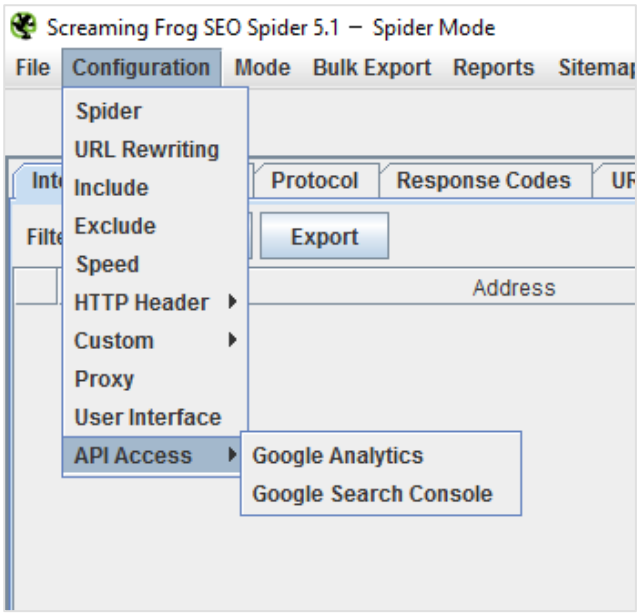

Fig. 2. Menu de connexion à Analytics et à la Search Console dans Screaming Frog Spider SEO.

2- Préparation des fichiers

Une fois le crawl terminé, nous allons exporter deux fichiers :

- La liste de toutes nos URL (ce que Gephi appellera des « nœuds ») ;
- La liste de tous les liens (ce que Gephi appellera des « edges »).

Pour les nœuds, il faut exporter le fichier internal_all (dans l'onglet « *Internal > Filter ALL > Export* ») :

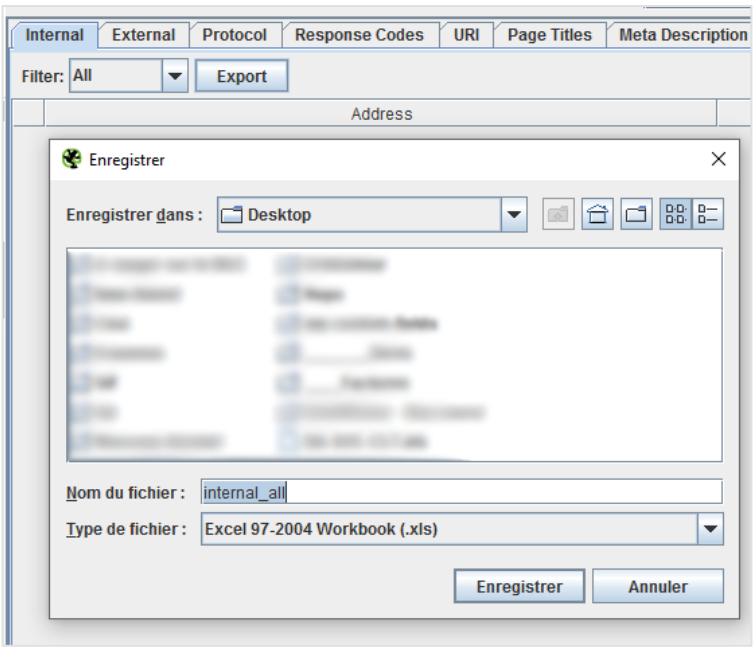

Fig. 3. Exportation des URL (nœuds).

Pour les liens, il faut exporter le fichier All_inlinks.

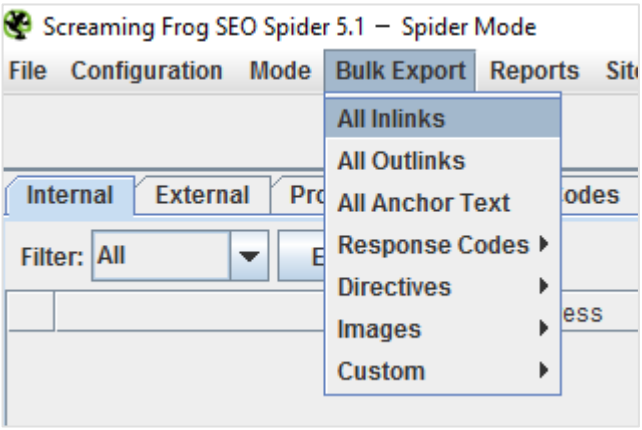

Fig. 4. Exportation des liens.

Ces fichiers sont au format Excel (xls). Il faut obligatoirement faire deux choses :

- Renommer certaines colonnes :
    - Pour les nœuds :
        - La colonne des URL doit avoir pour titre « ID » ;
        - La colonne affichant la balise Title pour les pages web (ou le Texte Alternatif pour les images) doit avoir pour titre « Label ».
    - Pour les liens :
        - La colonne de provenance du lien doit avoir pour titre « Source » ;
        - La colonne de destination du lien doit avoir pour titre « Target » ;
        - La colonne contenant l'ancre du lien doit avoir pour titre « Label ».
- les enregistrer au format CSV (séparateur point-virgule).

C'est d'ailleurs à cette étape que l'on peut en profiter pour y ajouter des données. Par exemple, on peut y inclure les données de Majestic SEO (Trust Flow, Citation Flow, nombre de domaines référents, Nombre de backlinks). Vous pouvez le faire quasiment automatiquement *via* SEO Tools ou en exportant une analyse depuis le site de Majestic SEO.

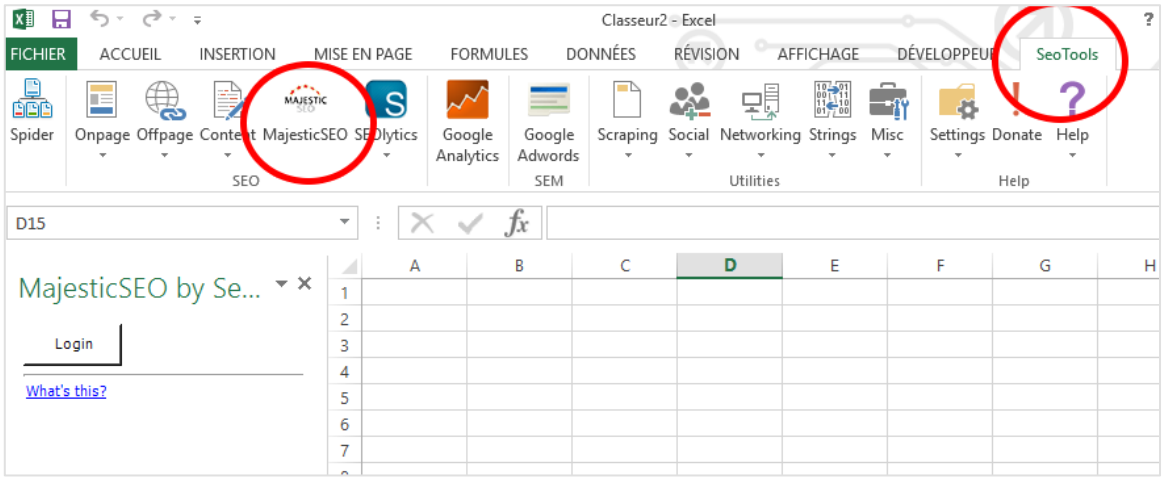

Fig. 5. Connexion Majestic SEO avec SEO Tools et Excel.

3- Importation des données

Lancez ensuite le logiciel Gephi. Celui-ci dispose de trois sections (trois onglets) :

- Le laboratoire de données (où l'on va importer nos données) ;
- La vue d'ensemble pour travailler nos visualisations ;
- La prévisualisation pour exporter nos graphiques.

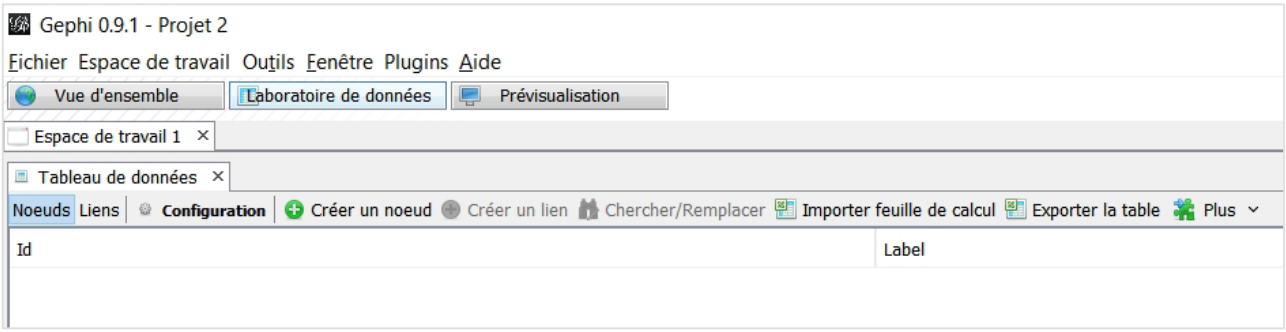

Fig. 6. Trois onglets permettent de travailler dans Gephi.

Dans le laboratoire de données :

1. cliquez sur « *Importer feuille de calcul* » ;
2. choisisez le fichier des nœuds ;
3. choisissez :
    a. le bon séparateur de votre fichier ;
    b. le bon encodage si nécessaire ;

c. pensez à préciser de quel type de fichier il s'agit (« En tant que table => Table des nœuds »).

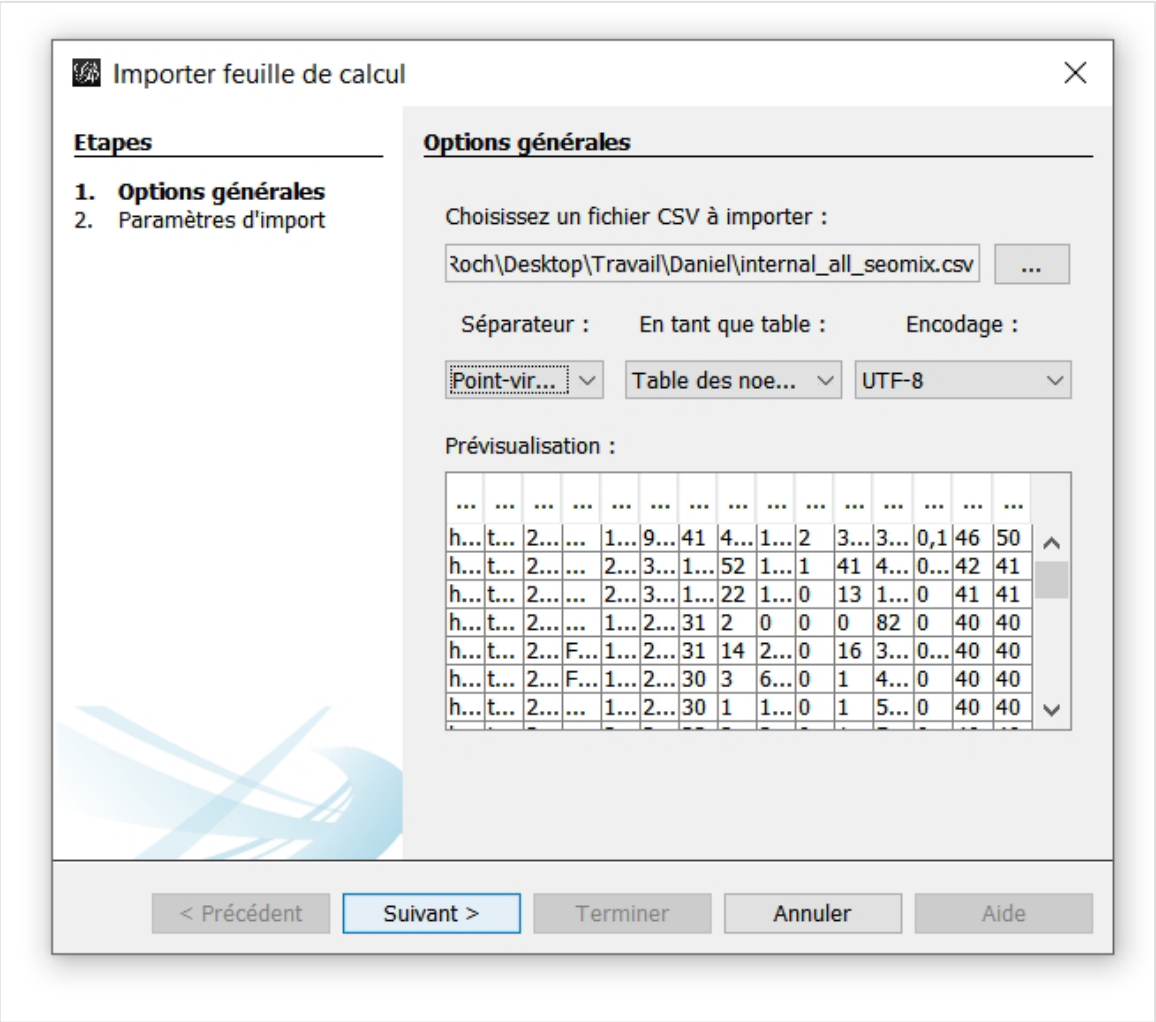

Fig. 7. Menu d'import de données de Gephi.

En cliquant sur « suivant », l'outil va vous lister toutes les colonnes qu'il a trouvées dans le fichier csv, et c'est à vous de lui indiquer ce dont il s'agit.

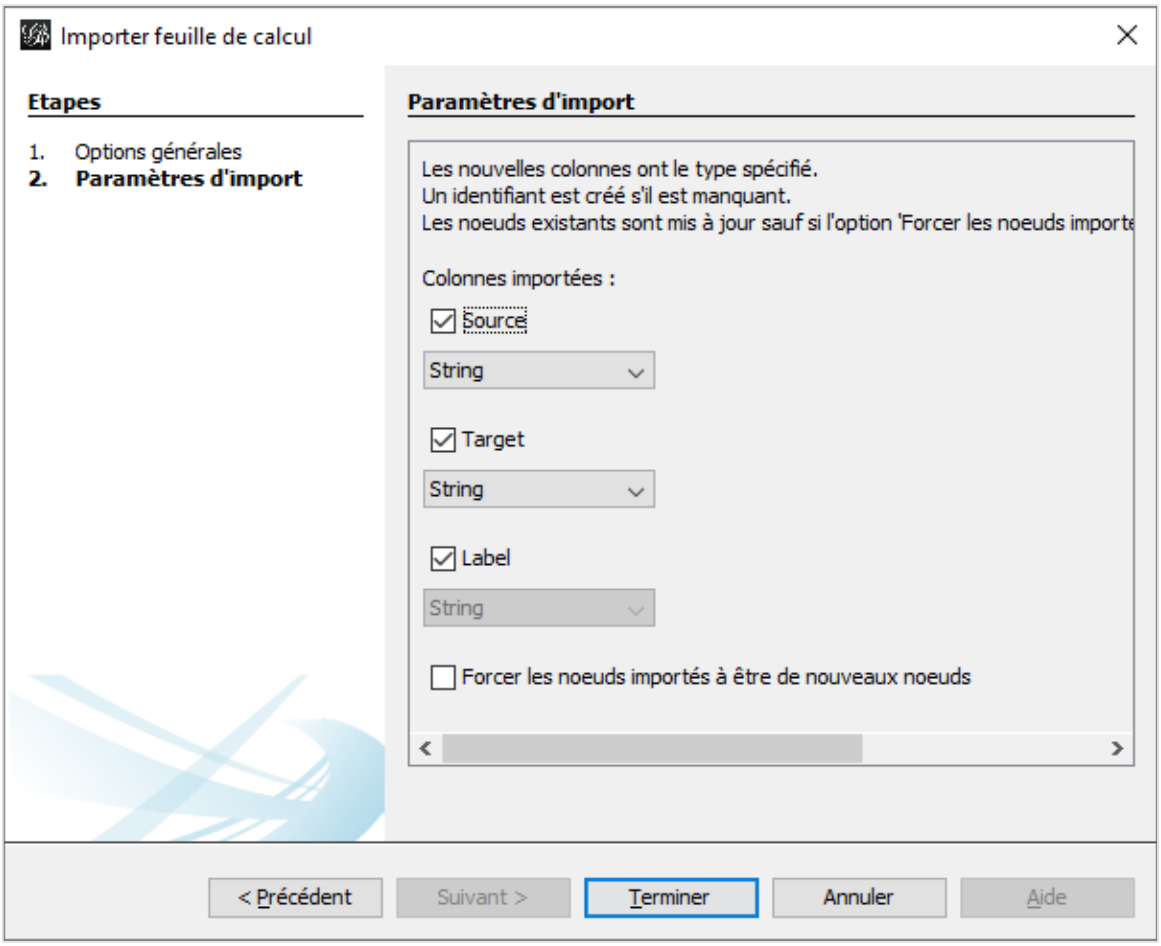

Fig. 8. Deuxième étape lors de l'import de données.

Dans cet écran, certaines listes déroulantes seront grisées : il s'agit des colonnes qui existent déjà (si vous aviez déjà importé des données) ou des colonnes obligatoires, c'est-à-dire celles que nous avions renommés au début. Pour les autres, vous devez obligatoirement indiquer ce dont il s'agit. Nous n'utiliserons que deux types de données :

- « String » pour tous les champs « Texte » ;
- « Integer » pour les chiffres.

Recommencez l'opération ensuite avec le fichier des liens. Quand vous importerez ce second fichier, ne cochez jamais « *Forcer les nœuds importés à être de nouveaux nœuds* » ni « *Créer les nœuds manquants* ».

Puis pensez déjà à sauvegarder...

4- Visualisation

En vous rendant dans le 1er onglet « *Vue d'ensemble* », vous allez pouvoir commencer à visualiser les données importées. Au départ, ce n'est pas utilisable car il vous faut trier le tout (voir figure 9).

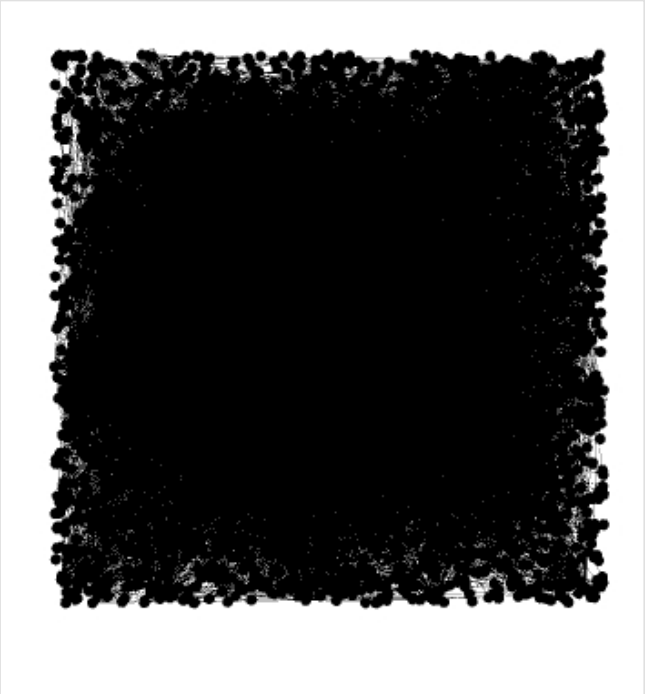

Fig. 9. Rendu par défaut de vos données. Y a du boulot...

Gephi affiche chacune de vos URL avec un rond, puis il trace des traits pour visualiser chaque lien entre les différends nœuds.

L'objectif est de pouvoir rendre visible ce graphique. Pour cela, nous devons utiliser les différentes options du logiciel :

- La modification des nœuds et des liens ;
- Les filtres ;
- Les calculs algorithmiques ;
- La visualisation spatiale.

La première étape est de hiérarchiser nos ronds. On peut le faire grâce à l'interface en haut à gauche « *Aspect* ».

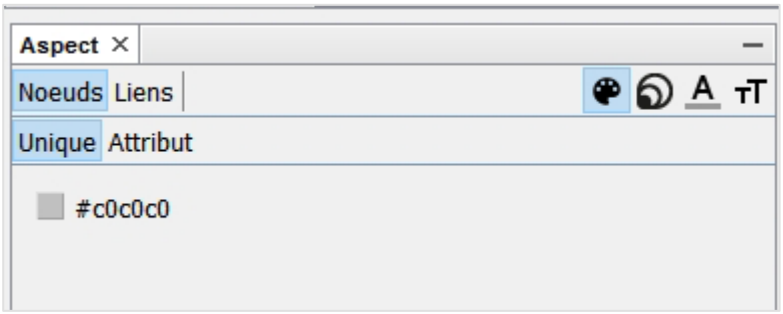

Fig. 10. Modification de votre visualisation commence par l'onglet « Aspect ».

Vous pouvez agir soit sur les Nœuds, soit sur les Liens, et à chaque fois avec :

- une modification basique « Unique » (par exemple pour tout colorer en gris) ;
- une modification selon un Attribut, c'est-à-dire selon une de vos données.

Vous pouvez modifier les différents éléments suivants, en utilisant les icônes en haut à droite du module « *Aspect* » :

- La couleur (comme sur l'image 10) ;
- La taille des nœuds ;
- La taille des textes ;
- La couleur des textes.

Sélectionnez donc le deuxième icône (taille des nœuds), puis sélectionnez « *Nœuds* > *Attribut* > *Degré* ». A vous ensuite de sélectionner une échelle cohérente. En général, une échelle 20-80 est idéale pour donner une taille suffisante à chaque rond tout en permettant de mettre en avant les nœuds qui reçoivent le plus de liens. Cela peut donc vous donner le type de rendu de la figure 11.

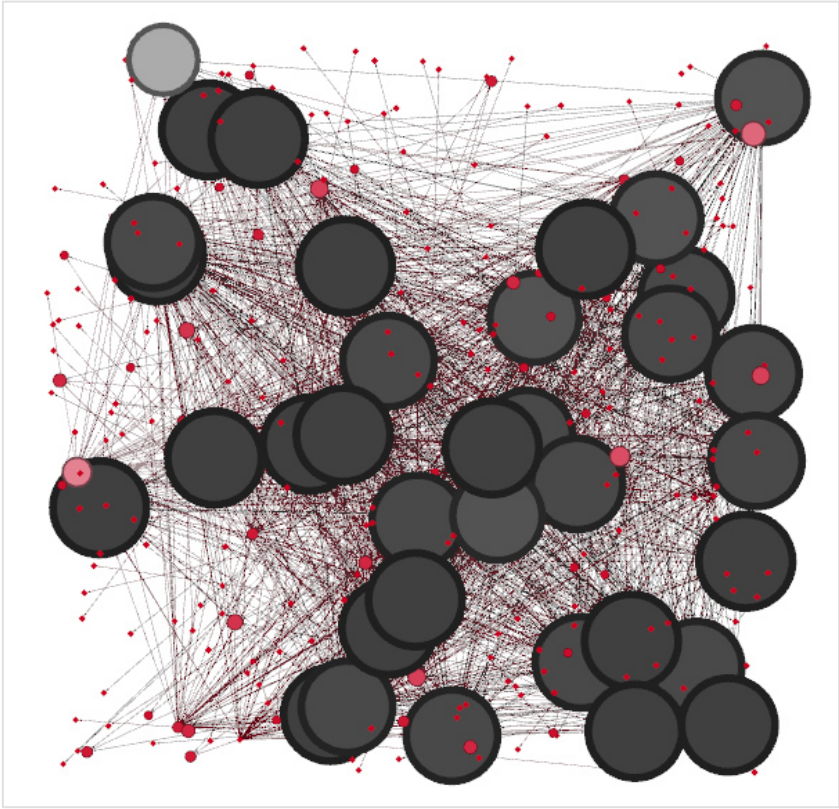

Fig. 11. Nous visualisons mieux les pages recevant beaucoup de liens.

Il faut ensuite séparer les différents nœuds. Pour cela, nous devons utiliser un autre élément de Gephi, la « *Visualisation* » en bas à droite.

Sélectionnez « *Force Atlas 2* » :

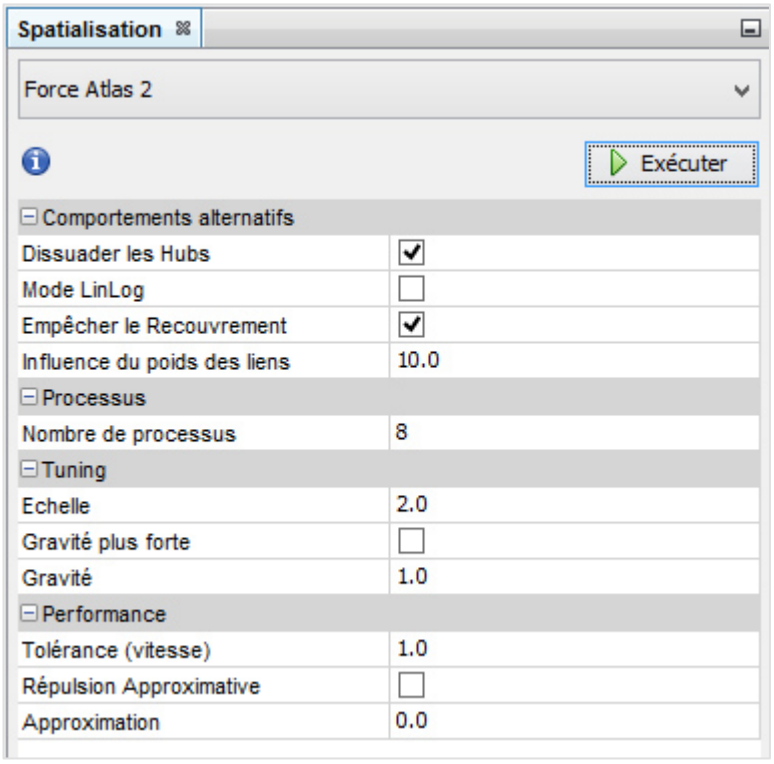

Fig. 12. Visualisation avec Force Atlas 2.

Cochez les cases « *Dissuader les hubs* » et « *empêcher le recouvrement* », puis cliquez sur « *Exécuter* ». En fonction du rendu, n'hésitez pas à modifier certains chiffres, notamment l'échelle, la tolérance ou encore l'influence du poids des liens.

Ce type d'algorithme de visualisation permet ainsi d'avoir le type de rendu de la figure 13.

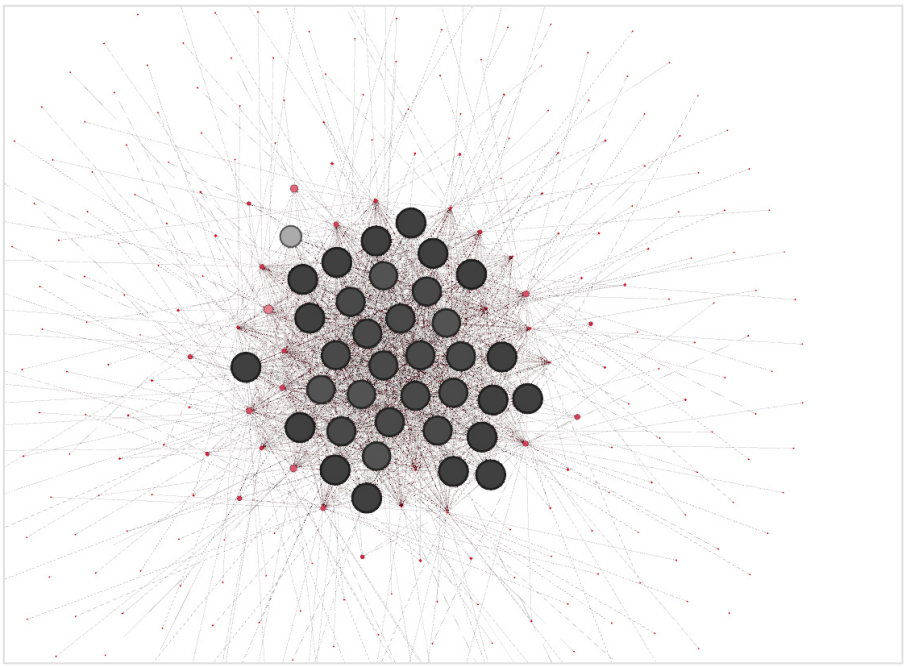

Fig. 13. Exemple de visualisation de données avec Force Atlas 2.

Sachez également avant d'aller plus loin que l'on peut faire des modifications manuelles. Prenons par exemple le fait de vouloir colorer une page précise (comme la page d'accueil). Quand vous sélectionnez un rond, vous pouvez faire un clic droit pour retrouver dans le tableau de données la ligne correspondante. Vous pouvez aussi faire l'inverse, à savoir faire un clic droit sur une URL dans le tableau de données pour la sélectionner dans la visualisation.

Fig. 14. Utilisez le dernier lien pour avoir des informations sur le rond sélectionné.

Une fois un nœud sélectionné, vous pouvez utiliser les boutons de l'interface principale. Dans la figure 15, on peut ainsi sélectionner :

- à gauche l'option couleur ;
- en haut la couleur et son opacité ;
- en haut à droite son comportement (ne colorer que le rond sélectionné ou les nœuds voisins également).

Fig. 15. Les options manuelles de modification

Et voici en figure 16 un exemple de ce que l'on peut faire en colorant manuellement certains nœuds en fonction de nos besoins.

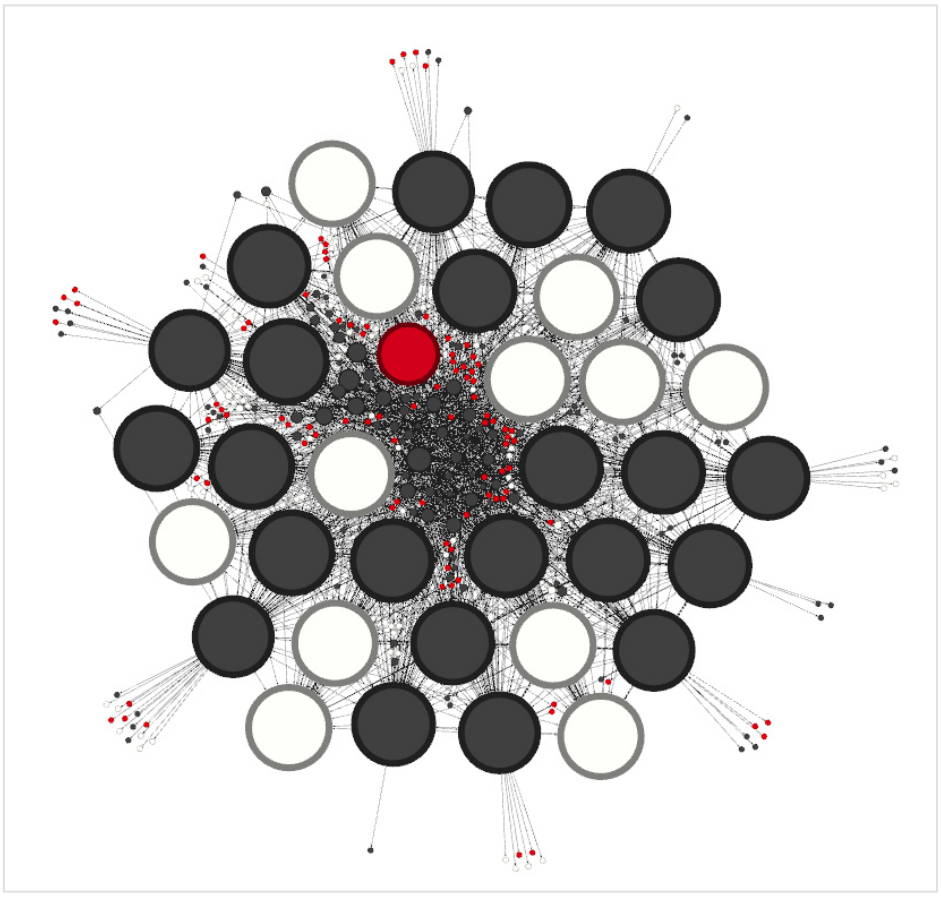

Fig. 16. En exemple de rendu visuel avec des modifications manuelles.

Sachez ensuite que l'on peut, avec la colonne de droite, utiliser certains algorithmes et des systèmes de filtrages. Nous en reparlerons dans les exemples à venir.

Et une fois de plus, pensez à sauvegarder dans un nouveau fichier votre travail.

Conclusion

Comme vous pouvez le constater, le rendu initial renvoyé par l'outil n'est pas spécialement utile. Nous le disions précédemment, Gephi est un excellent logiciel SEO, à condition de savoir ce que l'on veut analyser. Voici quelques pistes que l'on va détailler dans la seconde partie de cet article, le mois prochain :

- Visualiser les silos d'un site ;
- Visualiser le maillage mutuel des URL (A vers B et B vers A) ;
- Visualiser les pages populaires ;
- Faire ressortir les contenus inutiles ;
- Faire des tests SEO ;

- Visualiser les domaines référents de chaque URL ;
- Voir les URL avec peu de domaines référents.

En regardant la première partie de cet article, on se rend compte immédiatement de la complexité de ce logiciel. Pourtant, il pourra parfois vous donner une analyse bien plus fine et visuelle pour certaines actions de référencement, ou pour mesurer l'impact réel ou non d'une de vos modifications. Rendez-vous le mois prochain pour explorer plus en profondeur encore Gephi.

Daniel Roch, *Consultant WordPress, Référencement et Webmarketing chez SeoMix (http://www.seomix.fr/).*