

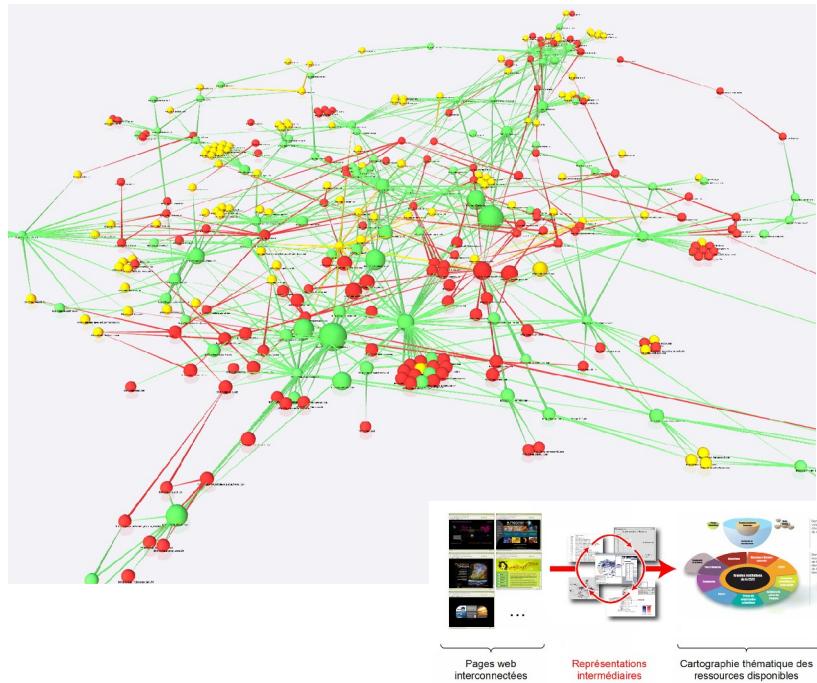
Fondation de la Maison des Sciences de l'Homme
Programme TIC-Migrations

**Méthodologies d'analyse
de corpus en Sciences Humaines
à l'aide du Navicrawler**

Rapport Final
Août 2007

Mathieu JACOMY (WebAtlas - CELSA - RITIMO, Paris)
Franck GHITALLA (WebAtlas - UTC, Nantes)

Sous la direction scientifique de Dana Diminescu (ENST, Paris)



Méthodologies d'analyse de corpus web en Sciences Humaines à l'aide du Navicrawler

Conforme à la version 1.5 du Navicrawler

Table des matières

Introduction.....	3
Contexte théorique : se repérer dans le web.....	4
La théorie des agrégats.....	4
Modèle en couches.....	4
Morphologie d'un domaine.....	6
Stratégie générale pour explorer un domaine.....	8
Recommandations pour bien utiliser le Navicrawler.....	10
Gérer la fermeture du corpus et la progression de la collecte.....	10
Principe de fermeture du corpus dans le Navicrawler.....	10
Faire le point et organiser son exploration.....	12
Maîtriser les libellés.....	16
Fonctionnalités de crawl : précautions méthodologiques.....	19
Le Navicrawler dans un écosystème logiciel.....	22
Exploiter les données avec un tableur.....	22
Sélectionner uniquement les sites incorporés.....	23
Sélectionner certains sites d'après un libellé.....	24
Trier d'après le nombre de pages visitées décroissant.....	24
Collecter une liste d'URLs avec le Navicrawler et Flem.....	25
Exploiter les données dans un logiciel de graphes.....	27
Voir un graphe issu du Navicrawler avec Guess.....	27
Modifier l'aspect visuel du graphe (introduction aux techniques).....	29
Visualiser les libellés.....	35
Utiliser des scripts pour Guess (étudier la distribution des liens).....	40
Interpréter la distribution des liens.....	47
Scénarios d'usage.....	50
Reconnaissance de terrain.....	50
Étudier le voisinage d'un site.....	56
Analyser la réponse d'un moteur de recherche.....	58
Détecter les centres.....	60
Circonscrire et analyser un domaine.....	65
Collecter et visualiser le pré-corpus.....	66
Interpréter la première carto.....	67
Etendre le corpus.....	68
Affinage et analyse de l'intérieur du corpus.....	69
Analyse de la frontière.....	70
Production de la carto finale.....	71
Mémo : récapitulatif des principaux conseils.....	72

Introduction

Ce document vous aidera à utiliser le Navicrawler efficacement. Il aborde différents aspects de sa mise en oeuvre, depuis les bases de la théorie du web jusqu'à la méthodo complète de l'exploration d'un domaine. Nous avons essayé d'être le plus « pratique » possible, ainsi ce document n'apporte pas de réponse à des questions précises sur le fonctionnement du Navicrawler ; en revanche il aborde les différentes situations auxquelles vous êtes confrontés lors de l'utilisation du Navicrawler et vous propose des exemples, des conseils et des explications, et des tutoriels pas-à-pas.

La première partie est consacrée à un rappel théorique nécessaire à la compréhension du web comme terrain d'analyse. En particulier les bases de la théorie des agrégats sont mobilisées pour une exploration cohérente du web en fournissant des repères utiles dans l'expérience de navigation comme dans la progression méthodologique.

La seconde partie regroupe différents conseils pour utiliser efficacement les fonctions de circonscription du corpus, les libellés et le crawl. Vous y apprendrez comment déterminer un terrain cohérent sans « patauger » dans le web, et comment organiser efficacement vos descripteurs. La fonction crawl est souvent utilisée à mauvais escient : vous comprendrez le cadre dans lequel le crawl peut vous aider, les limites de cette fonctionnalité et les usages à éviter.

La troisième partie est dédiée à l'exploitation des données issues du Navicrawler dans d'autres logiciel. Vous y trouverez un tutoriel pour l'utiliser avec un tableur et l'extension Flem, et un long tutoriel sur l'analyse de corpus sous la forme de graphes, exemples à l'appui. Le logiciel de graphes que nous avons choisi de mobiliser est Guess et comme les autres outils peut être téléchargé librement.

Enfin vous trouverez dans la quatrième partie plusieurs scénarios d'usages que vous pourrez suivre pas à pas pour accomplir différentes tâches telles que l'analyse des résultats d'un moteur de recherche ou l'étude du voisinage d'un site donné. L'accent est mis ici sur la dimension méthodologique de l'activité, notamment la circonscription du terrain et l'interprétation des données. A la fin de cette partie vous trouverez le mise à plat de l'exploration complète d'un domaine, sous la forme d'un organigramme d'aide à la décision.

Contexte théorique : se repérer dans le web

La théorie des agrégats

La théorie des agrégats, issue de travaux statistiques probabilistes sur l'étendue et la structure du web, stipule que les documents qui traitent du même sujet ont une plus forte probabilité d'être connectés (par des liens hypertextes) : « **Qui se ressemble se connecte** ». Les sites d'un même domaine sont souvent connectés, et forment ainsi ce qu'on appelle un « **agrégat** », centré sur une thématique. Tout le web n'est pas constitué d'agrégats, toutefois leur existence permet la détection de motifs sémantiques par un moyen non sémantique. On cherche les agrégats là où le web est le plus dense, et on vérifie dans un second temps à quels discours correspondent les ressources. Les moteurs de recherches exploitent ce principe : ils peuvent calculer le degré d'importance d'un site sans prendre en compte ses mots. C'est grâce à ça qu'ils vont si vite !

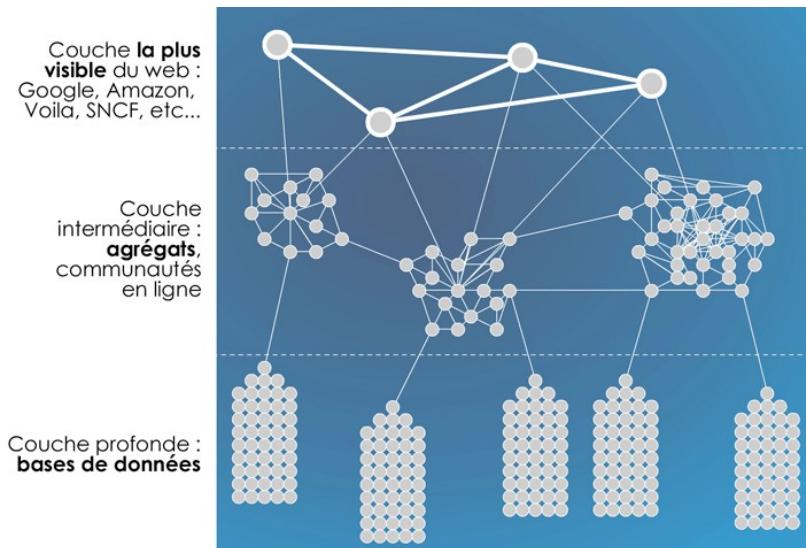
En réalité il existe toutes sortes d'agrégats, dont certains sont même si peu denses qu'on ne devrait même pas les appeler ainsi. Nous n'en savons pas beaucoup plus à l'heure actuelle, mais nous pouvons explorer le web à la main (sans utiliser les moteurs de recherches) pour observer comment sont reliés les sites. Ainsi nous parlerons de « **domaine** » pour désigner l'ensemble des ressources qui parlent d'une même chose, quelle qu'elle soit. Ex. : *le domaine de la broderie est constitué en agrégat. Le domaine des nanotechnologies n'est pas constitué en agrégat.*

Modèle en couches

Lorsqu'on explore un domaine, on retombe en permanence sur certains sites qui nous paraissent donc très importants. Or si une partie de ces sites sont les plus « célèbres » du domaine, la plus grande partie est trop générique pour appartenir au domaine. Il s'agit de sites qui reviennent dans tous les domaines, comme Wikipédia, Sourceforge, Spip, Adobe, Microsoft, Google... Il peut s'agir aussi d'instances nationales ou continentales qui ne sont pertinentes qu'à grande échelle : la SNCF, le gouvernement de la France, la commission européenne, le CNRS, les ministères... Enfin il peut s'agir des sites majeurs de grands domaines voisins, dont la notoriété empiète sur la pertinence : ainsi du site de la Fondation Nicolas Hulot qui apparaît dans le corpus des Parcs Naturels (le lien est légitime mais le site est tout de même hors-sujet).

Les grands sites génériques reçoivent tellement de liens qu'ils sont présents largement hors des domaines qui font leur spécificité. Ainsi des sites ministériels qui sont si connectés entre eux qu'ils sont tous présents ensemble dans de

nombreuses explorations. Tous ces sites fortement connectés forment ce qu'on appelle « la couche haute du web ». La couche qui est la plus pertinente lorsqu'on explore une thématique est la couche dite « intermédiaire » : il faut éliminer la couche haute qui concentre une grande partie des liens pour voir apparaître la couche intermédiaire. Enfin il existe une couche « basse » qui contient essentiellement des bases de données et se trouve à de nombreux clics de la couche haute. Elle est très spécialisée et est restée longtemps inaccessible aux moteurs de recherches, d'où sa dénomination initiale de « deep web » ou « invisible web ».



Lorsqu'on explore un domaine, il y a deux façons de sortir de la zone pertinente. La première façon consiste à entrer dans la couche haute. Les sites sont trop génériques, très connectés et alourdiraient considérablement le corpus si on devait les prendre en compte. La seconde façon consiste à passer de la thématique actuelle à une thématique connexe, d'un agrégat à un agrégat voisin. Les sites sont alors en rapport parfois explicite avec les sites pertinents mais ils sont néanmoins hors-sujet. Par exemple si l'on explore les sites sur les poissons, on trouvera des sites sur les algues. Le passage d'une thématique à une autre se fait souvent graduellement : on trouve souvent des thématiques qu'il faut choisir subjectivement de garder ou non. Dans notre exemple il s'agit de la communauté des aquariophiles, qui s'intéressent tout à la fois aux poissons (d'aquarium) et aux algues (d'aquariums). Ces sites de thématiques voisines sont délicats à gérer, parce qu'ils posent en permanence la question du contour du domaine que l'on explore.

Il arrive que la thématique choisie ne soit pas dotée d'un agrégat dédié. Dans le meilleur des cas, l'agrégat n'est pas encore constitué : les sites ne sont pas ou peu connexes. Autre cas épineux, la thématique n'existe qu'au sein d'une thématique plus large : des sites plus génériques surviennent en permanence et relient les sites pertinents entre eux. Il est alors nécessaire d'élargir un peu le champ de l'investigation. Enfin il se peut que la thématique soit très

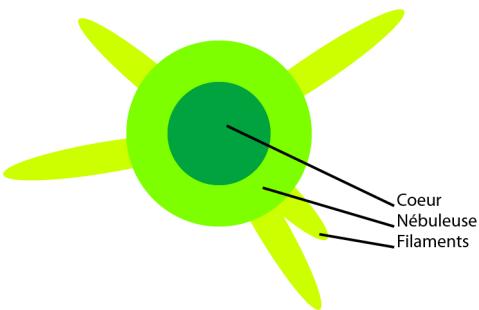
transversale. C'est notamment le cas des explorations par localité géographique : les sites sont souvent reliés par thème de contenu mais rarement par localité géographique. Dans cet exemple, lorsque la localité géographique entre dans le contenu il y a agrégation (collectivités territoriales, patrimoine local) mais en dehors de ça les sites se fragmentent en autant de thématiques. L'exploration est alors très fastidieuse car elle oblige à sans cesse sortir du champ d'investigation pour y re-reentrer afin d'obtenir des données suffisamment complexes.

Les agrégats facilitent grandement l'exploration car grâce à eux, une fois entré dans la thématique on maximise les chances d'y rester pendant la navigation. Lorsqu'on ne bénéficie plus de cet effet pour une raison ou une autre, la trop grande quantité d'information non-pertinente risque d'enliser l'exploration. Cet enlisement se produit aussi naturellement dès que la majeure partie de l'agrégat a été collectée : il faut alors mettre un terme à l'exploration.

Morphologie d'un domaine

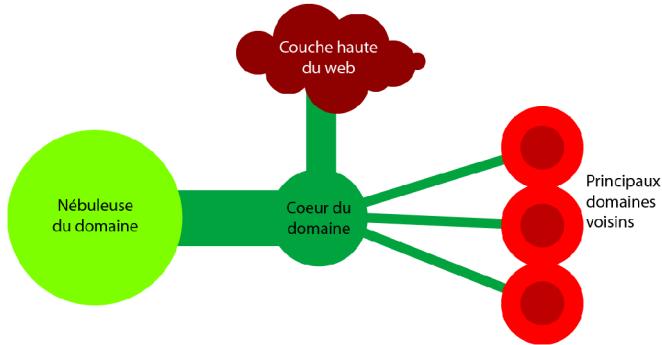
On distingue trois composantes à un domaine ou agrégat pour se repérer :

- Le **coeur**, qui contient des sites souvent très gros et très fortement connectés
- La **nébuleuse**, qui comprend la majeure partie des sites du domaine mais pas les plus connectés
- Les **filaments**, qui est constituée de sites souvent petits, qui entrent dans le champ de la thématique mais qui sont peu connectés avec le domaine lui-même.



Le cœur est entouré de la nébuleuse, à laquelle se raccrochent les filaments. Selon le domaine étudié, la taille de ces composantes peut varier. Le cœur est l'élément le plus identifiable tandis que les filaments sont les plus difficiles à récupérer lors de l'exploration. Les graphes des corpus terminés présentent souvent cette forme identifiable. Lorsqu'on explore un domaine il est très utile de savoir se situer dans cette géographie primaire ; en effet les trois zones ne

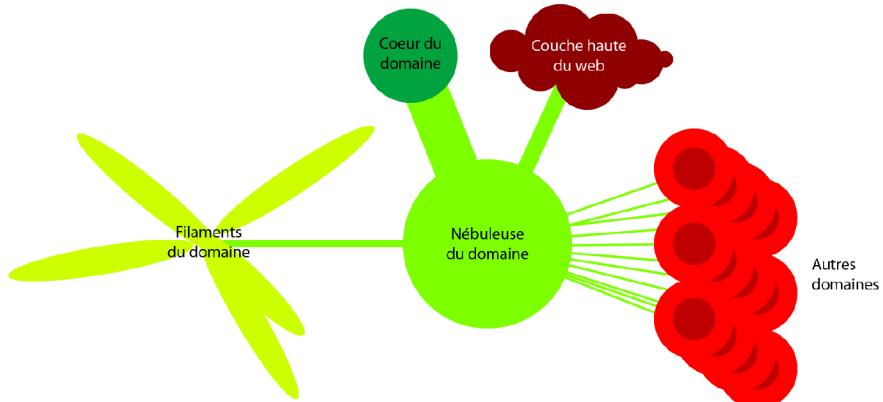
sont pas reliées de la même façon au reste du web. Nous allons détailler ces caractéristiques en commençant par le cœur.



Le cœur est essentiellement connecté avec la nébuleuse du domaine. Mais comme nous l'avons vu dans le modèle en couches, il est également connecté à la couche haute du web. En outre il est lié dans une moindre mesure avec d'autres domaines. Repérer le cœur est une priorité parce qu'il permet d'accéder dans de bonnes conditions à la nébuleuse. Les sites du cœur sont reconnaissables à ceci que :

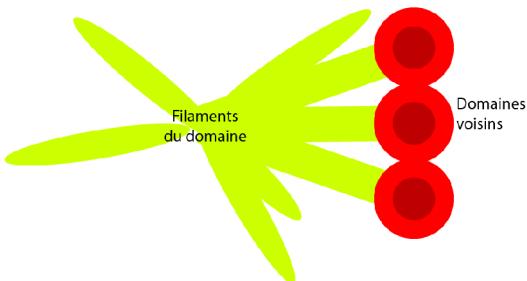
- Ils sont explicitement dédiés au domaine
- Ils disposent de nombreux liens dont on peut évaluer la pertinence élevée à vue d'oeil
- Ils comprennent souvent un contenu riche, voire font office d'arbitres pour le domaine
- Ils sont fortement connectés avec les couches supérieures et notamment les institutions, la recherche ou les grandes entreprises.
- Les sites du cœur reviennent sans cesse dans les liens des autres sites

Explorer le cœur suppose que l'on est capable d'éliminer la couche haute du corpus (nous verrons comment plus loin). Remarquons qu'il n'est pas toujours évident de rester dans le cœur en naviguant, mais qu'en revanche il est facile d'y re-reentrer à partir de la nébuleuse.



La nébuleuse est principalement connectée au cœur du domaine, ainsi que dans une moindre mesure à la couche haute du web. Elle est aussi connectée aux filaments et à des domaines voisins. Attention, là où le cœur est connecté avec les principaux domaines voisins, la nébuleuse est connectée à une multitude de domaines voisins mais souvent très faiblement. C'est pour cette raisons que nous l'appelons nébuleuse : il n'est pas toujours facile de passer d'un site de la nébuleuse à un autre sans passer par le cœur, car la densité de liens est faible. Cette situation peut s'expliquer par le fait que la nébuleuse contient beaucoup plus de sites que le cœur, mais que ces sites sont moins centrés sur le domaine. Dans la nébuleuse peuvent se trouver des mini-agrégats, qui correspondent aux sous-domaines de la thématique explorée, et la frontière entre le cœur et la nébuleuse n'est pas nette du tout. Cependant il est souvent plus délicat de savoir si un site est ou non dans la thématique lorsqu'on est dans la nébuleuse, car l'information est moins travaillée que dans le cœur.

Comme la nébuleuse est fortement connectée au cœur, il est facile de retrouver le cœur à partir d'elle. Il suffit de naviguer et de repérer quels sont les sites qui reviennent souvent dans les liens : ce sont les sites du cœur.



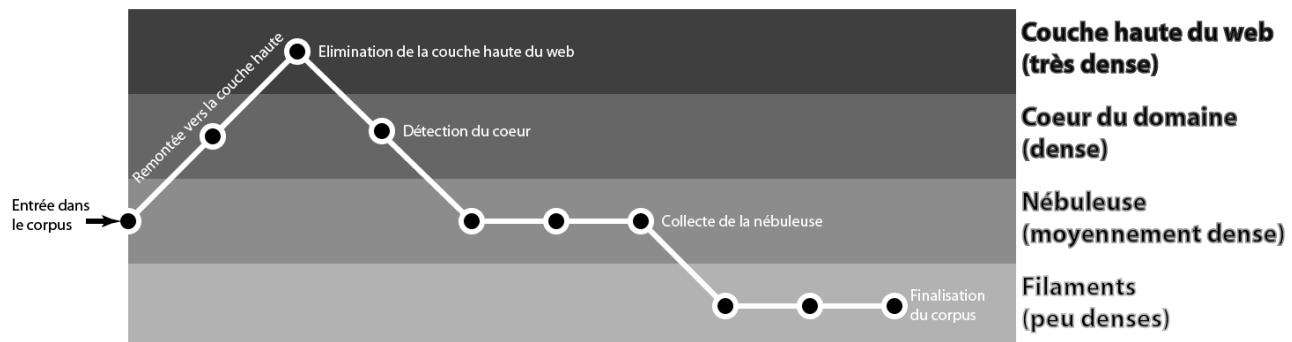
Enfin, les filaments sont situés en bordure de la nébuleuse et sont essentiellement connectés avec des domaines voisins. De ce fait ils sont très pénibles pour l'exploration, car un site d'un filament est quasiment en dehors du domaine excepté que le contenu du site nous force à le prendre en compte. En revanche les filaments pointent dans de nombreux cas vers des domaines voisins du domaine exploré, de sorte qu'ils sont un relativement bon indicateur du contexte de la thématique dans le web – encore que ceci reste à prouver dans le cas général.

Stratégie générale pour explorer un domaine

Comme nous venons de le voir avec le modèle en couche et la morphologie d'un domaine, plus on est bas dans les couches – c'est-à-dire le moins le tissu hypertextuel est dense – plus suivre les liens nous amènera soit dans les couches supérieures soit à l'extérieur du domaine. Autrement dit les liens nous dirigent par nature vers les zones plus denses, jusqu'à la couche haute du web, le « centre de tous les centres ». Ce mécanisme est naturel puisque par

définition les zones denses sont celles qui sont les plus liées, donc celles où statistiquement nous avons le plus de chance de nous diriger.

Nous prônons l'exploration « **du haut vers le bas** », des couches denses vers les couches peu denses. Intuitivement, parce qu'on ne peut pas lutter contre ce courant qui nous ramène vers les couches denses. Mais en partant du haut vers le bas, on élimine pas à pas les couches supérieures pour faire apparaître les couches plus profondes. Si nous partions du bas vers le haut, il y aurait d'emblée tellement de sites des couches hautes que le tri serait impossible, sans compter qu'il faudrait sans cesse « redescendre » pour ne pas oublier des sites. L'exploration consiste donc à partir d'un site du cœur ou de la nébuleuse, et à se « laisser porter par le courant » vers les sites plus importants jusqu'à ce qu'on arrive aux premiers sites trop génériques. A ce stade il s'agit d'éliminer ces sites (la couche haute du web) pour laisser le cœur apparaître. Une fois le cœur collecté, la nébuleuse apparaît. La collecte de la nébuleuse suppose d'éliminer des sites qui sont au même niveau de densité mais hors-domaine, il s'agit de la phase d'expansion du corpus. Enfin l'exploration de filaments est possible et permet de tester que toute la nébuleuse a bien été collectée. Les filaments ne sont pas systématiquement collectés, car le travail est souvent trop lourd pour un faible gain.



Cette stratégie a le mérite de proposer une lecture méthodique des tâches à accomplir. En effet les couches peu denses n'apparaissent qu'à condition que les couches supérieures aient été traitées ou éliminées. Par contre cette stratégie nécessite des outils appropriés. L'outil que nous utilisons à cette fin est le Navicrawler, qui permet de traiter ou d'éliminer les couches hautes. Nous allons maintenant voir ce principe en détail.

Recommandations pour bien utiliser le Navicrawler

Gérer la fermeture du corpus et la progression de la collecte

La documentation du Navicrawler explique en détail le processus de fermeture d'un corpus en milieu ouvert. Nous allons d'abord le rappeler ici pour l'approfondir.

Principe de fermeture du corpus dans le Navicrawler

Vous naviguez sur le web avec le Navicrawler pour constituer un corpus. Naturellement vous disposez d'une fonctionnalité qui vous permet de choisir si vous voulez garder un site pour votre corpus ou si vous n'en voulez pas :

- Pour choisir si vous voulez garder ou rejeter un site, une page de ce site doit être ouverte dans la fenêtre principale de Firefox.
- Lorsque vous accédez à une page d'un site que vous n'aviez pas navigué jusqu'ici, **il est gardé dans le corpus par défaut**.
- Si vous ne voulez pas du site ouvert dans votre corpus, cliquez sur le bouton « **refuser** ».
- Les sites que vous gardez sont appelés « **sites incorporés** » et les sites que vous refusez sont appelés « **sites écartés** ».

En outre, le Navicrawler vous aide en répertoriant les liens hypertextes des pages que vous visitez. Non seulement il les détecte et vous les propose sous forme de liste (cliquez sur « liens dans la page ») mais il les classe selon que les liens pointent vers d'autres pages du *même* site ou vers des pages *d'autres* sites. Le Navicrawler détecte donc des sites que vous n'avez pas visités, vers lesquels il y a des liens depuis les sites que vous avez visités. Ces sites sont appelés « **sites prochains** ».

Tous les sites que connaît le Navicrawler sont soit incorporés, soit écartés, soit prochains. Un corpus constitué par le Navicrawler est composé de ces trois types de sites et des liens entre eux. Selon ce que vous projetez de faire avec votre corpus, il peut être pertinent ou non de conserver ces trois types de sites ou de ne garder, par exemple, que les sites incorporés (cochez les cases correspondantes au moment de l'export). Les sites écartés sont intéressants pour étudier ce qui constitue le voisinage de votre thématique. Les sites prochains représentent la portion du voisinage que vous n'avez pas étudiée.

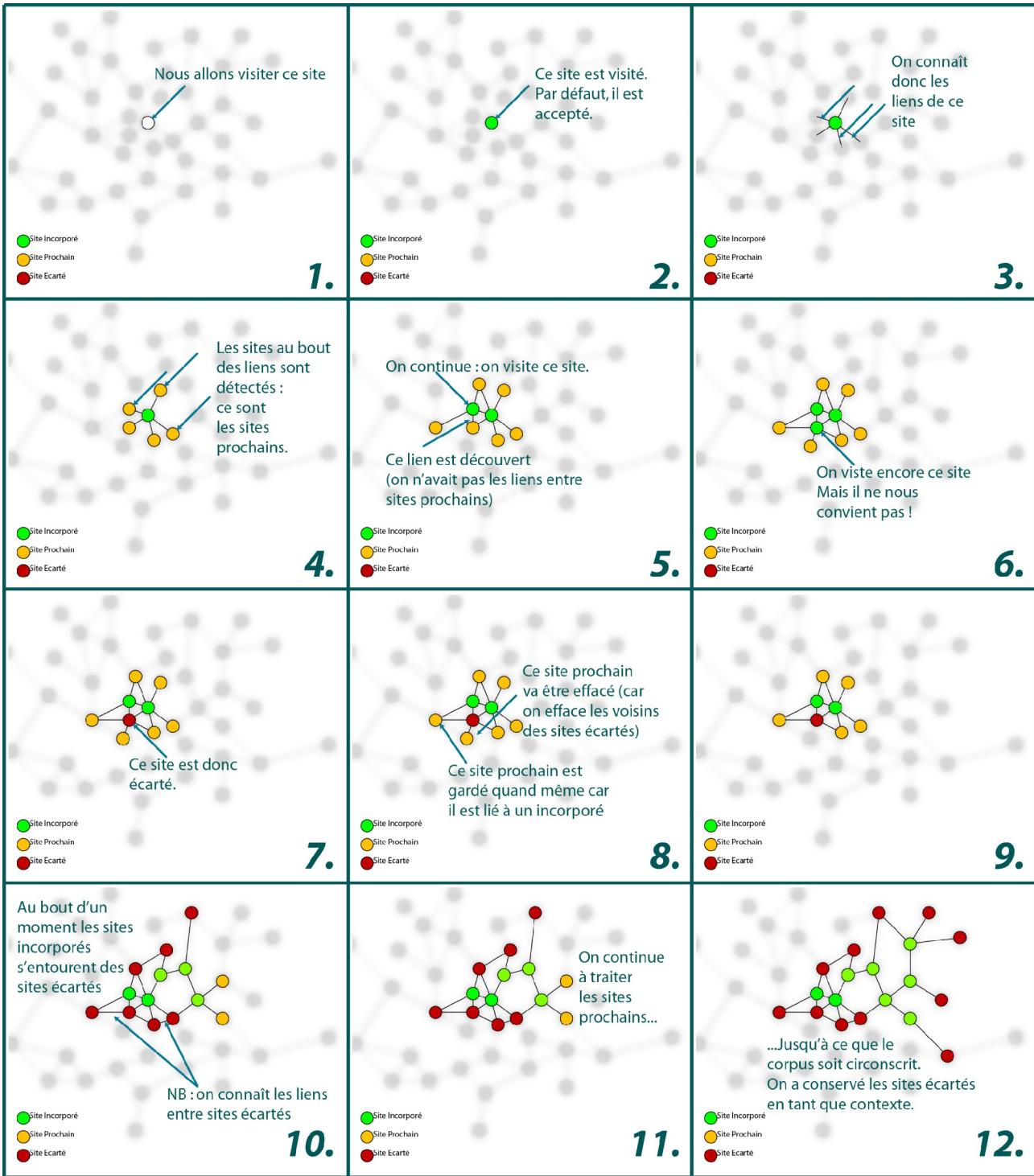
- Les sites incorporés sont traditionnellement colorés en **vert**, les sites écartés en **rouge** et les sites prochains en **orange**.
- Lorsque vous naviguez d'un site A à un site B, en suivant les liens

hypertextes, le site B a été détecté comme prochain lorsque vous étiez sur le site A. Donc la chronologie pour un site est généralement d'être prochain, puis incorporé, puis s'il le faut écarté. Les sites prochains sont ceux que vous visiterez potentiellement dans la suite de l'exploration, d'où leur nom. Prochain rappelle aussi « proche » : ce sont les sites qui sont proches des sites que vous avez incorporés.

- Les sites écartés tirent leur nom du fait que vous les avez refusés, mais aussi du fait qu'ils sont à l'écart de votre corpus sans pour autant être ignorés. Ils peuvent être pris en compte dans l'analyse de votre corpus.
- **Les sites écartés n'ont pas de sites prochains** : nous considérons que lorsque vous avez refusé un site, il n'est pas nécessaire de vous proposer son voisinage. Ceci est très important dans le mécanisme d'exploration propre au Navicrawler. En effet, si vous classez les sites prochains les uns après les autres, au bout d'un moment vous n'aurez plus que des sites incorporés et des sites écartés.

La logique du Navicrawler est de vous inciter à étendre votre exploration à partir des sites prochains. En théorie, vous pouvez vous contenter de prendre la liste des sites prochains et de les classer jusqu'à ce qu'il n'y en ait plus. En pratique, ceci ne marche que pour les petits corpus, car à chaque fois que vous incorporez un nouveau site, ses voisins sont ajoutés aux sites prochains (mais pas lorsque vous écartez). Le risque est donc d'avoir tellement de sites prochains que l'exploration s'en trouve empêchée. Nous verrons plus loin comment nous en sortir.

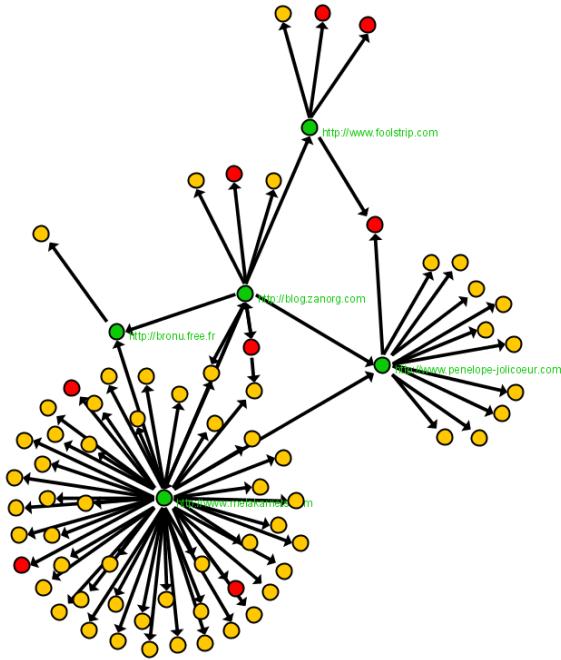
Le schéma suivant illustre la façon dont s'articulent les sites incorporés, écartés et prochains au cours de l'exploration.



Faire le point et organiser son exploration

Il est possible de visualiser l'exploration en cours avec un logiciel de graphes dans le but de faire le bilan de l'avancement. Le graphe ci-dessous est produit suivant les mêmes conventions graphiques que les illustrations ci-dessus. Le début d'une exploration des blogs BD a été exporté en GDF, et ouvert dans Guess. La position des noeuds est donnée empiriquement par l'algorithme GEM inclus dans GUESS. Le nom a été affiché seulement pour les sites incorporés. On peut remarquer que seulement 5 sites ont été incorporés et 8

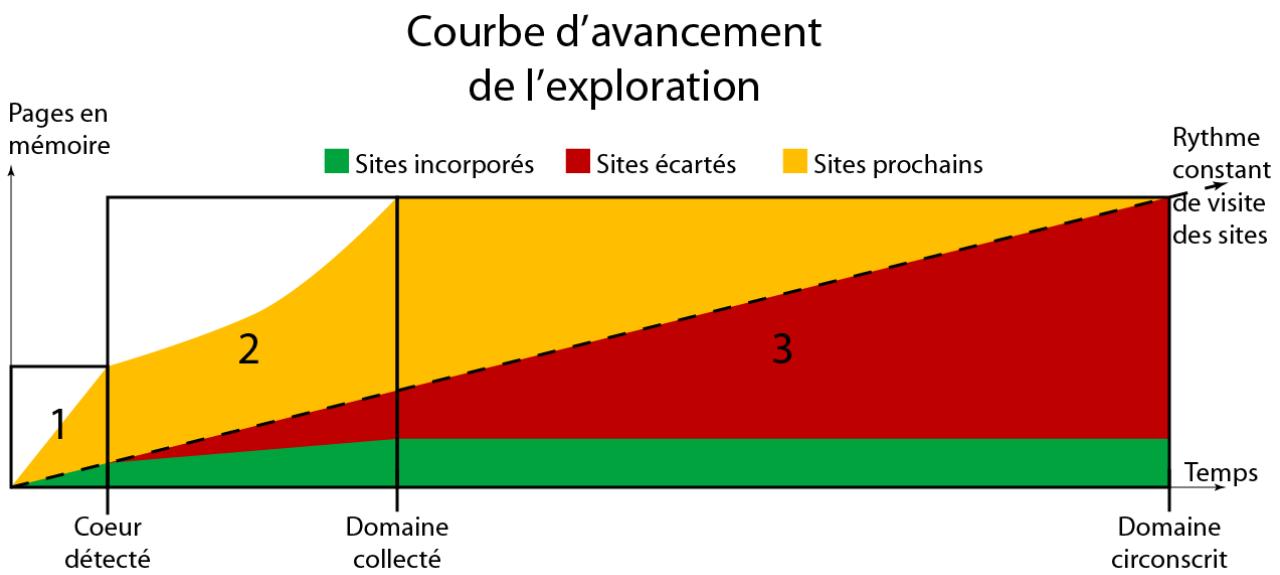
écartés. Il y a un grand nombre de sites prochains : c'est très souvent le cas. Le site « melakarnets » en vert en bas à gauche référence la majorité de ces sites à explorer. Il est probable qu'il fasse plus ou moins office de portail, ou en tout cas de relais majeur vers le domaine. Une bonne option serait alors d'explorer méthodiquement les sites cités par « melakarnets » pour orienter la suite de l'exploration.



Le voisinage de certains sites doit donc être épluché systématiquement. Le choix des sites sur lesquels on effectue ce « zoom » est au choix de l'explorateur, et de l'approche qu'il adopte dans son travail. Le cas le plus fréquent se présente lorsque le coeur a été détecté. Dans ces circonstances, il est utile de zoomer sur ces sites pour traiter tout leur voisinage : ces sites sont riches en liens, et ce sont les plus pertinents puisqu'ils constituent le coeur. Pour ce faire il est d'abord nécessaire de collecter un maximum de pages sur lesquelles se trouvent des liens. La fonction de crawl peut être mobilisée à cet effet mais il suffit la plupart du temps de traquer les pages de liens et les pages de partenaires. Une fois ceci fait, la liste des sites référencés par le site du coeur où l'on zoome est proposée par le Navicrawler dans l'encart « **sites cités** ». Cliquer sur cet encart affiche la liste, qui présente des sites dans la couleur correspondant à leur statut. Le travail d'épluchage se résume à visiter tous les sites prochains présents dans cette liste. Cliquez sur un site de la liste avec le bouton « Ctrl » enfoncé vous permettra de l'ouvrir dans un nouvel onglet. Ainsi, la liste des sites cités reste disponible lorsque vous affichez l'onglet initial. Remarquons que lors de l'exploration, l'incorporation de nouveaux sites

augmente souvent le nombre de sites prochains. C'est bien sûr le cas ici mais aucun site prochain ne sera ajouté à la liste de sites cités par le site zoomé, ce qui permet de gérer l'avancée du travail plus facilement.

L'augmentation de sites prochains devient un problème dès qu'on se donne pour objectif de tous les classer et que le corpus dépasse les quelques dizaines de sites incorporés. Il est classique d'obtenir dix fois plus de sites prochains que de sites incorporés, et ce dès le début de l'exploration. De la même façon, le corpus une fois circonscrit sera entouré de beaucoup plus de sites écartés qu'il n'y aura de sites incorporés. Or, la circonscription complète du domaine nécessite de classer tous les sites prochains, c'est-à-dire que la majeure partie de l'exploration sera consacrée à écarter des sites à la chaîne, pour peu de nouvelles incorporations. Voici comment se déroule une exploration classique :



- Lors de la phase 1, la plupart des sites visités sont incorporés. En effet, en partant du principe que les sites de départ sont bien choisis, la plupart de leurs voisins sont également pertinents. A l'issue de la phase 1, la principale partie du cœur du domaine a été visitée. Une grande quantité de sites prochains sont détectés car les sites du cœur sont très connectés.
- La phase 2 est l'expansion du corpus, du cœur vers la nébuleuse. Les sites autour de ceux précédemment incorporés sont triés, certains sont incorporés à leur tour tandis que d'autres sont écartés (les sites de la couche haute du web et les sites de domaines voisins). La quantité de sites prochains augmente irrégulièrement. NB : il arrive qu'au début la quantité de sites prochains n'augmente pas très vite, parce que les sites aux alentours du cœur ont beaucoup de liens en commun (ils sont bien centrés sur le domaine) ; au contraire, à la périphérie du domaine il y a beaucoup de « bruit » qui se traduit par une grande quantité de sites prochains qui seront

par la suite écartés.

- La phase 3 est le temps de la clôture (ou circonscription) du corpus. Quasiment tous les sites correspondant aux critères ont été incorporés, et de ce fait il n'y a quasiment pas d'augmentation de la quantité de sites prochains. Au contraire, tous les sites prochains sont peu à peu écartés jusqu'à ce qu'il n'y en ait plus. Cette phase est mécaniquement la plus longue, sans pour autant être la plus nécessaire. Elle permet toutefois de s'assurer que le corpus est le plus exhaustif possible.

Dans de nombreux cas il n'est pas nécessaire de mener la circonscription jusqu'au bout, sachant que lorsque le domaine est étendu la charge de travail devient tout simplement trop importante pour ce type d'outil.

Le choix du moment où arrêter l'exploration relève d'un arbitrage entre la qualité et l'efficacité. Plus on arrête tôt plus il est probable que des ressources importantes auront été oubliées, mais plus on arrête tard moins il y a de sites pertinents dans ceux qu'il reste à classer. On cherche donc à arrêter assez tôt mais avec un maximum de garanties. Dans le cas où le domaine se constitue en agrégat, il est possible d'obtenir cette garantie assez vite pourachever l'exploration dans de bonnes conditions.

- S'il est possible d'identifier un **coeur** au domaine et si **tous les sous-domaines** sont représentés dans le corpus
- Lorsque le cœur ainsi que les sites majeurs de chaque sous-domaine ont été **visités en profondeur** (il faut que les liens sortants de ces sites soient bien collectés)
- Alors on peut **éliminer des sites prochains** tous ceux qui ne sont pas cités par au moins deux sites différents. On néglige ainsi les sites qui sont cités par un seul ou aucun site majeur. Cette fonction est disponible dans l'onglet « utils » du Navicrawler.
- Les sites prochains restants doivent être classés avec attention, car ils sont probablement soit des sites importants du domaine soit des sites de la couche haute. une fois ceci fait, il est possible d'arrêter l'exploration avec la garantie qu'**aucun site majeur n'aura été oublié**.

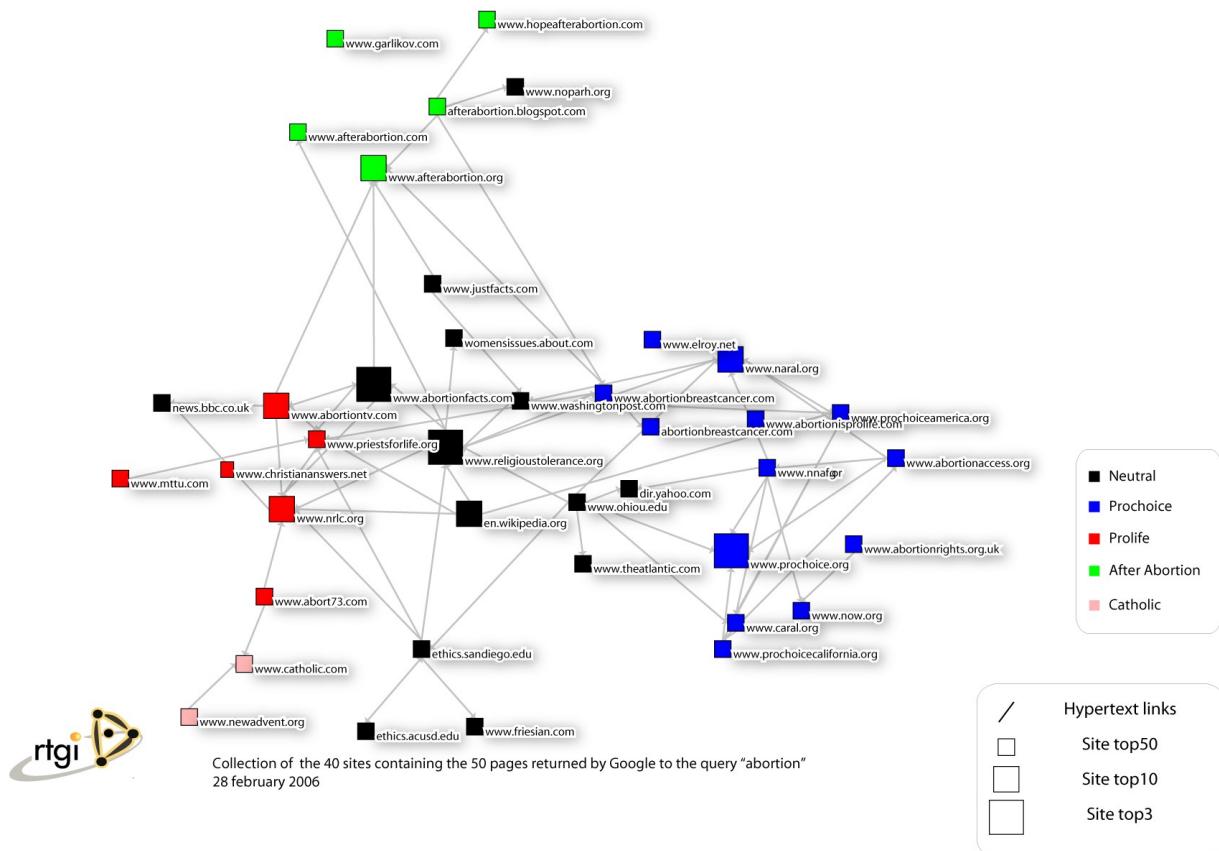
Cette technique permet d'accélérer la phase d'expansion du corpus, lorsqu'on se trouve dans la nébuleuse du domaine. Elle permet d'évacuer d'un coup de nombreux petits sites hors-sujet. Le nombre de sites prochains est divisé par au moins 2 dans la plupart des cas, parfois même par 10. Par contre, cette technique n'est pas efficace pour éliminer les sites de la couche haute.

L'élimination des sites de la couche haute peut s'avérer pénible. L'expérience est ici bonne conseillère puisque ces sites sont connus et reviennent dans de tous les domaines : il est possible de les identifier rien qu'à leur URL. Voici donc comment procéder. Affichez la liste des sites prochains de la session, ou la liste

des sites cités par le site courant. Si vous reconnaisssez un site à éliminer dans la liste, appuyez sur la touche « Majuscule » de votre clavier, puis sans la relâcher cliquez sur le site à éliminer dans la liste. Il sera alors écarté instantanément.

Maîtriser les libellés

Les libellés du Navicrawler doivent être compris au regard de la tâche de cartographie. Ce système permet de rajouter de l'information « classifiante » aux sites (les noeuds du graphe) de façon à la projeter graphiquement sur le graphe, sous la forme de couleurs ou d'autres paramètres graphiques. Le cas typique consiste à choisir un code couleur pour un ensemble de libellés afin de faciliter la compréhension d'une carte-graphe. Voici un exemple de code-couleur appliqué à des libellés, sur une étude des 50 premiers résultats de Google pour la requête « abortion » en anglais :



Deux systèmes co-existent dans le Navicrawler : les libellés indépendants et les groupes de libellés. Un libellé indépendant servira par exemple à décrire si le site est personnel ou collectif : on crée le libellé « personnel » et on choisit pour chaque site si c'est « oui », « non » (site collectif) ou « report » (indécidable). Un groupe de libellés servira par exemple à choisir la raison sociale d'un site entre plusieurs libellés possibles : « association », « entreprise », « institution » etc. La principale différence tient au fait que pour les libellés indépendants on choisira une couleur pour « oui », une autre pour « non », une autre pour « reporté » et

une dernière pour « non-classé » ; tandis que pour les groupes de libellés on choisira une couleur par libellé du groupe, plus une couleur pour « reporté », plus une couleur pour « non-classé ».

Le problème principal dans l'utilisation des libellés est qu'on ne sait pas *a priori* quels descripteurs seront pertinents pour la session. Les libellés sont donc créés au cours de l'exploration. A l'heure actuelle le Navicrawler ne dispose pas de fonction avancées de gestion des libellés. Néanmoins la possibilité de choisir « reporté » permet de cadrer un minimum la tâche de labélisation. En effet il est courant de trouver des « cas-limites » pour un libellé donné. Il est alors nécessaire d'enregistrer le fait qu'un site soit un cas limite, car il arrive que les cas limites deviennent si nombreux que ça en devient l'indice d'un mauvais libellé, soit qu'il soit « décalé » par rapport à la réalité du terrain, soit qu'il soit insuffisant et doive être complété d'un autre libellé. La fonction « **reporté** » permet donc de notifier qu'il a été choisi de ne pas choisir, ce qui diffère tout-à-fait des sites non-classés qui incarnent simplement ce qu'il reste à labéliser. Au demeurant les cas-limites sont souvent intéressants et il n'y a pas nécessité de trouver systématiquement une case où ranger tous les sites... Notez que le système reporté/non-classé s'applique aux libellés indépendants comme aux groupes de libellés.

La différence entre 10 libellés indépendants et un groupe de 10 libellés tient au fait qu'ils soient ou non exclusifs les uns des autres. Dans un groupe de libellés, chaque site est « oui » pour un et un seul libellé et « non » pour tous les autres du groupe, à moins que le groupe soit en « reporté » ou « non-classé ». Voici quelques exemples de groupes de libellés :

- « Langue » : Francophone, Anglophone, Hispanophone, Autre langue, Multilingue.
- « Géolocalisation » : Paris-IDF, Province, DOM-TOM, Etranger, Inconnu.
- « Raison sociale » : Association, Entreprise, Institution, Site personnel, Site-vitrine, Autre.

Dans ces exemples chaque site correspondra à un et un seul de ces libellés, ou bien c'est un cas-limite. Ainsi un site sera soit francophone, soit hispanophone etc. Ces groupes sont des **couches de description** qui sont souvent les mêmes pour les différents corpus mais qui doivent toujours être adaptées. Les trois grandes couches de description qui reviennent souvent sont :

- La **raison sociale**
- La **thématische** (le contenu)
- La **géolocalisation**

Cependant le contenu de ces groupes, c'est-à-dire les libellés qu'on choisit d'y inscrire, sont déterminés par l'angle d'approche de l'utilisateur et par les

caractéristiques du domaine. Selon qu'on s'intéresse uniquement au web francophone ou pas, selon le degré de généralité du domaine auquel on s'intéresse, selon les limites que l'on a donné à son corpus a priori, les libellés pertinents ne sont pas du tout les mêmes. Parfois le système de groupe n'est pas assez libre et des libellés indépendants restent plus efficaces. Un exemple typique est l'étude de sites migrants : dans ce cas il est plus utile de garder des libellés indépendants pour chaque langue à cause de la multiplicité de sites multilingues, et de créer en supplément un groupe « multilinguisme » contenant les libellés exclusifs suivants : « monolingue », « traduction séparées », « coexistence des langues ».

Les libellés indépendants peuvent être utilisés comme descripteurs classiques ou comme descripteurs méthodologiques pour s'aider à cadrer l'exploration. Le libellé « à réinvestir » est par exemple utile pour garder en mémoire certains sites qu'il serait intéressant de garder pour d'autres explorations dans d'autres domaines. De même le libellé « très générique » aidera à gérer les sites dont on a du mal à déterminer s'ils appartiennent au cœur du domaine ou à la couche haute. Voici maintenant quelques exemples de libellés indépendants « classiques » :

- Propose la vente en ligne
- Dispose d'un forum
- Flux RSS
- Portée internationale
- Site amateur

Une erreur courante dans l'utilisation des libellés réside dans la confusion entre la tâche de **sélection** (ou circonscription) du corpus et la tâche de **description**. Ces deux tâches doivent rester hermétiquement séparées. La raison en est simple, c'est justement que la sélection et la description des ressources sont interdépendantes. Selon ce qui est sélectionné, les critères de description ne seront pas les mêmes (effet « angle d'approche »). Inversement c'est en décrivant qu'on s'aide à sélectionner les ressources de façon cohérente (processus de circonscription du terrain). Dans l'idéal il s'agit donc d'alterner sélection et description jusqu'à stabiliser un terrain proprement circonscrit, puis d'effectuer la dernière étape de description. Si l'on effectue les deux tâches en même temps, l'effet « angle d'approche » joue contre la compréhension du terrain et reporte sans cesse l'étape de circonscription. Le noeud du problème réside dans le fait qu'un corpus en milieu ouvert (le web) ne se décrit qu'au regard de son contexte. Il faut donc se donner une limite puis étudier simultanément le corpus et son contexte pour les comprendre ensemble ; c'est d'ailleurs la raison pour laquelle conserver les sites « écartés ». Une description rigoureuse ne se construit qu'au regard de la limite qu'on se donne pour

circonscrire le terrain. Pourtant les libellés peuvent être utilisés pour comprendre le terrain avant de le circonscrire : dans ce cas il est recommandé de recommencer les libellés après la circonscription.

Fonctionnalités de crawl : précautions méthodologiques

Le Navicrawler n'est pas un bon crawler, il se situe à la lisière de la navigation et du crawl et permet de comprendre comment fonctionne un véritable crawler. Au delà de quelques centaines de sites, le Navicrawler n'arrive plus à gérer la masse d'information ; tandis qu'un crawler autonome sait gérer des centaines de milliers de pages, voire des millions de sites (et même bien plus dans le cadre des moteurs de recherche).

Le Navicrawler peut naviguer « tout seul », et ce faisant il collecte l'information de la même façon que lorsque vous naviguez manuellement : de là naît sa fonction de crawler. Autrement dit le Navicrawler distingue les deux composantes d'un crawler : le comportement de navigation et la collecte. C'est ainsi qu'il vous propose par défaut de collecter comme un crawler mais sur la base de votre propre navigation. Votre navigation est bien plus experte que les règles de comportement basiques d'un crawler. Bien sûr elle est moins rapide, mais elle vous permet de comprendre les pages sur lesquelles vous naviguez et c'est tout l'intérêt de cet outil. N'oublions pas que les cartes que le Navicrawler aide à construire ne peuvent pas être pertinentes si les données sources (le web) ne sont pas interprétées en amont. La première règle est donc : **n'utilisez le crawl qu'avec parcimonie.**

Vous pouvez utiliser les fonctionnalités de crawl à condition de bien cadrer leur utilisation. Si vous savez ce que vous faites, le Navicrawler vous permettra d'aller plus vite et plus loin dans votre exploration. Mais si vous ne faites pas attention, vous vous retrouverez avec un corpus où seront mélangés des sites que vous avez incorporés à la main et des sites que le Navicrawler aura incorporés de lui-même, c'est-à-dire sans analyse. De fait, un tel corpus ne peut pas être interprété correctement. La seconde règle est donc : **ne laissez pas le crawler incorporer des sites lorsque vous en avez déjà incorporés manuellement.**

Pour ce faire, il suffit de paramétriser le crawl en distance 0. En effet la distance représente le nombre de sauts entre sites successifs que vous autorisez le crawler à effectuer. Si vous choisissez une distance 0, vous n'autorisez pas le crawler à effectuer le moindre saut à l'extérieur des sites points d'entrée. Ainsi, il se contentera de rester dans ces sites et d'indexer leurs pages.

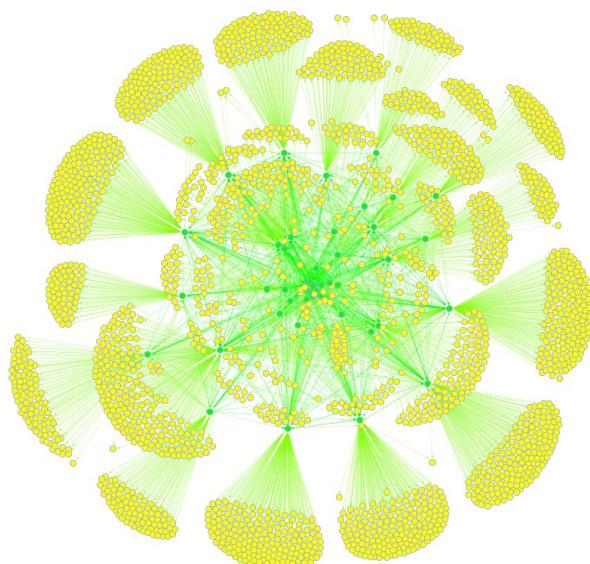
Crawler à l'intérieur d'un seul site est intéressant, car ceci vous permet de visiter de nombreuses pages et de collecter ainsi la plupart des liens vers des sites prochains. Deux remarques cependant. D'abord, il n'est pas toujours

souhaitable d'avoir tous ces liens. En effet certains sites ont des liens pertinents dans leur page de liens, mais peuvent en receler un grand nombre de gênants dans les pages plus profondes. Typiquement, lorsque le site héberge un forum. Ensuite, il est souvent plus efficace d'aller directement à la page de liens ou de partenaires pour collecter les liens principaux : nous vous recommandons de faire ainsi plutôt que d'utiliser le crawl, ce sera plus rapide. Mais si vous souhaitez pour une raison ou une autre collecter de nombreuses pages du site, utilisez un crawl en distance 0. Il faut alors choisir la bonne profondeur. La profondeur est le nombre de « clics » qu'il faut pour atteindre une page à partir de la page de départ. **La plupart du temps, en distance 0, la bonne profondeur de crawl est 2.** La plupart des pages se situent à cette profondeur, à une exception cependant : lorsque le point de départ est la page d'accueil et que cette page d'accueil n'est qu'une petite animation ou un logo, alors vous avez intérêt à passer en profondeur 3 pour le même résultat. Afin d'avoir un point de repère, dites-vous que le nombre de pages à visiter est multiplié par 10 à chaque niveau de profondeur. Ainsi une profondeur 4 dans un site qui dispose d'un tel niveau de profondeur représente déjà 10000 pages à visiter : c'est trop. Cependant si le site vous semble particulièrement peu dense (pages peu nombreuses mais aussi peu liées), vous pouvez envisager une telle profondeur. C'est à vous d'évaluer la situation. Mais gardez à l'esprit que le Navicrawler n'est pas adapté à la collecte de gros sites comme Wikipedia ou TF1 par exemple, qui disposent de plus de 20 niveaux de profondeur. Dans tous les cas nous vous conseillons de faire un petit test en profondeur 1 pour vous donner un ordre de grandeur de ce que vous demanderez au crawler.

Vous pouvez aussi utiliser le Navicrawler avec une distance de 1 ou plus, seulement donc si vous n'avez pas déjà incorporé de sites manuellement. De même que pour la profondeur, vous pouvez calculer approximativement le nombre de sites qui seront incorporés par le crawler en multipliant par 10 pour chaque distance. Ainsi une distance de 3 donne 1000 sites incorporés et 10000 sites prochains : encore une fois c'est trop. Il faut aussi considérer que pour chaque site, le nombre de pages visitées dépend de la profondeur de crawl comme expliqué ci-dessus. La principale utilisation du crawl par sites, donc avec une distance de 1 ou plus, est de dégrossir le travail d'exploration tout au début. À partir d'un ou plusieurs points d'entrée, il s'agit alors d'obtenir rapidement une bonne liste de sites pour étudier de quoi elle se compose. Dans ce cas l'idéal est d'effectuer un crawl en distance 1 et profondeur 1, puis de recommencer avec d'autres sites au besoin. Il arrive parfois qu'on s'intéresse à un domaine particulièrement peu dense, dans ce cas des crawls en distance 2 ou plus se justifient. Ce cas est cependant très rare, vous vous en rendrez compte si un crawl en distance 1 ne donne qu'un tout petit nombre de sites incorporés et de sites prochains. Mais la plupart du temps, n'utilisez pas une distance supérieure à 1.

Si vous lancez malgré tout dans un crawl d'envergure, prévoyez du temps pour que le Navicrawler travaille : lancez le crawl le soir et vous verrez le lendemain matin si votre ordinateur disposait d'assez de mémoire vive... Vous pouvez accélérer le crawl en désactivant les images et le Flash : les pages se téléchargeront plus vite. Des extensions dédiées comme PrefBar¹ vous permettront de le faire très simplement.

A titre d'exemple, nous avons effectué un crawl en profondeur 1 et en distance 1 à partir d'un skyblog banal, avec les images et le Flash désactivés, 10 onglets de crawl et 32ms de délai. Le crawl a duré 7 minutes, et donne 51 sites incorporés et 2640 sites prochains. La distance 2 n'est pas envisageable, car elle donnerait 2700 sites incorporés pour plusieurs dizaines de milliers de sites prochains. Une profondeur 2 resterait envisageable par contre, à condition de multiplier le temps nécessaire par 10 ou plus. Voici le graphe produit :



Rappelez-vous que le crawler ne sait pas distinguer les couches du web. Il ne sait pas lorsqu'il entre dans la couche haute, or c'est justement ce que nous voulons éviter. Vous remarquerez dans le graphe ci-dessus que 10 à 20 sites incorporés (en vert) sont responsables de la majeure partie des sites prochains (en jaune). Chaque « champignon » jaune manifeste l'ensemble des sites prochains cités par un unique site incorporé (le site vert à la base du « champignon »). La plus grande partie des sites jaunes se compose d'une bonne quinzaine de ces champignons, ce qui signifie que seule une petite partie des sites incorporés est à l'origine de la plupart des sites prochains détectés. Ces sites appartiennent vraisemblablement à la couche haute du web. Il serait alors judicieux de visiter ces sites, de ne les garder que si c'est nécessaire, et recommencer le crawl uniquement à partir des sites pertinents. Cette méthode « pas à pas » permettra de diminuer drastiquement le « bruit » (sites non-pertinents) et le temps nécessaire à la collecte.

¹ <http://prefbar.mozdev.org/>

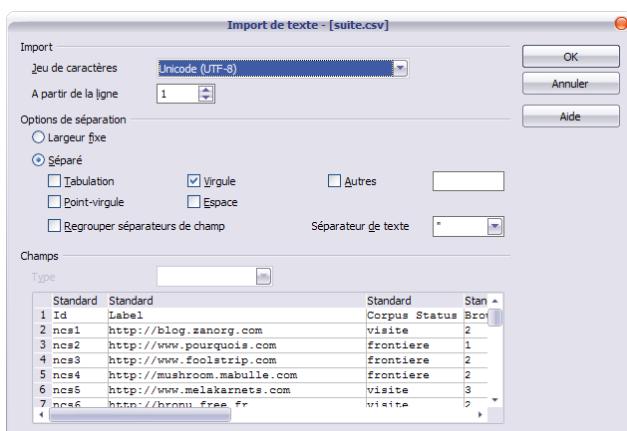
Le Navicrawler dans un écosystème logiciel

D'autres outils sont utiles et parfois nécessaires pour traiter les données issues du Navicrawler. Nous allons évoquer ici une autre extension pour Firefox, *Flem*², ainsi que deux logiciels libres qui permettent de manipuler ces données. Le premier logiciel est *OpenOffice Calc*³, un tableur bien connu, qui pourrait aussi bien être remplacé par *Excel*. Le second est un logiciel de visualisation de graphes appelé *Guess*⁴. Ces outils sont répertoriés sur le portail web-mining.fr dans la section « outils », ainsi que d'autres logiciels utiles pour le traitement des données⁵.

Remarque préliminaire : le format WXSF du Navicrawler n'est utile que pour sauvegarder ses sessions et les rouvrir plus tard. Les autres logiciels ne connaissent pas ce format de fichier propre au Navicrawler, et le Navicrawler ne peut pas rouvrir d'autre format de fichier.

Exploiter les données avec un tableur

La façon la plus simple de visualiser ses données consiste en l'utilisation d'un tableur, outil bien connu sur lequel nous ne nous étendrons pas. Depuis la version 1.5 le Navicrawler génère des fichiers CSV, compatibles avec les tableurs ; ces fichiers sont exportés par paire : un fichier pour les liens et un fichier pour les sites. Le plus important est le CSV des sites. Ouvrons-le avec OpenOffice Calc. Le logiciel demande quel format d'import utiliser ; en général les options par défaut sont satisfaisantes : encodage UTF-8, séparation par virgule, et séparateur de texte par guillemet.



Le fichier se présente donc sous la forme d'un tableau, dont chaque ligne correspond à un site. En colonne vous trouverez respectivement :

- L'identifiant du site : pas très intéressant.

² Télécharger Flem : [http://www.web-mining.fr/flem_\(firefox_links_explorer_module\)](http://www.web-mining.fr/flem_(firefox_links_explorer_module))

³ Télécharger OpenOffice : <http://fr.openoffice.org/>

⁴ Télécharger Guess : <http://graphexploration.cond.org/>

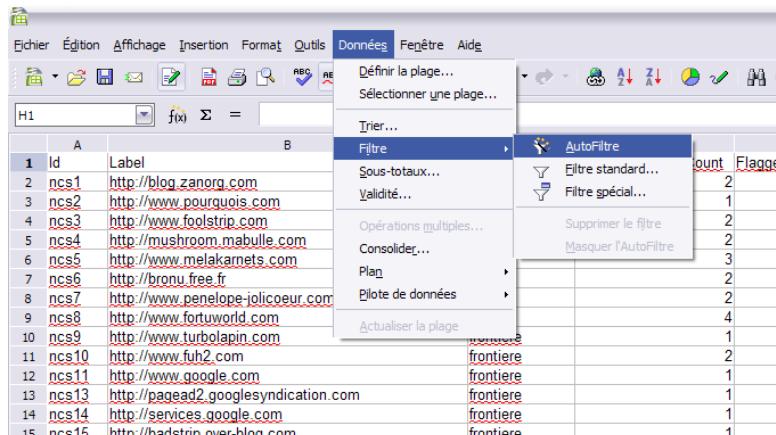
⁵ <http://www.web-mining.fr/outils>

- Le nom du domaine !
- L'état : incorporé, prochain, écarté. Attention, la dénomination peut être l'ancienne (visité=incorporé, voisin=prochain, frontière=écarté).
- Le nombre de pages visitées.
- Le nombre de pages marquées.
- La type du noeud : si vous n'utilisez pas d'heuristiques, ce sont tous des sites !
- Les libellés indépendants (valeurs possibles : oui, non, report, non-classé)
- Les groupes de libellés (valeurs possibles : les libellés du groupes, report, non-classé)
- Si vous utilisez des heuristiques, d'autres propriétés peuvent apparaître.

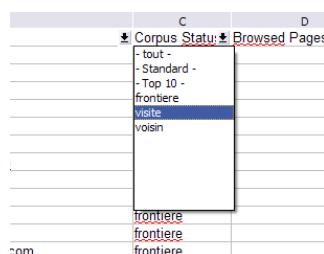
A partir de là, les principales fonctions utiles dans le tableur sont le **tri** et la **sélection**. Le tri fonctionne sur les valeurs numériques essentiellement, et la sélection sur les valeurs textuelles. Voici quelques exemples de traitements intéressants.

Sélectionner uniquement les sites incorporés

Dans le menu « Données », sélectionnez « Filtre » puis « Autofiltre ».



Des petits menus déroulants vont s'ajouter en haut de chaque colonne. Dans la troisième colonne, cliquez sur ce menu et sélectionnez « incorporés » ou « visité ».



Seuls les sites incorporés apparaissent maintenant. Pour tout afficher de nouveau, sélectionnez « tout » dans le même menu.

Sélectionner certains sites d'après un libellé

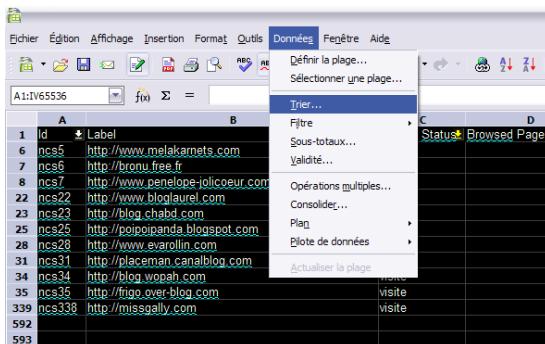
De la même façon vous pouvez n'afficher que les sites correspondant à une certaine valeur pour un libellé. Dans l'exemple suivant (blogs BD) nous allons sélectionner les sites où le libellé « dessin amateur » vaut « non » :

F	G	H
Type	TAG Dessin_amateur	
- tout		
- Standard -		
- Top 10 -		
non		
non-classe		
oui		
report		
non		
non		

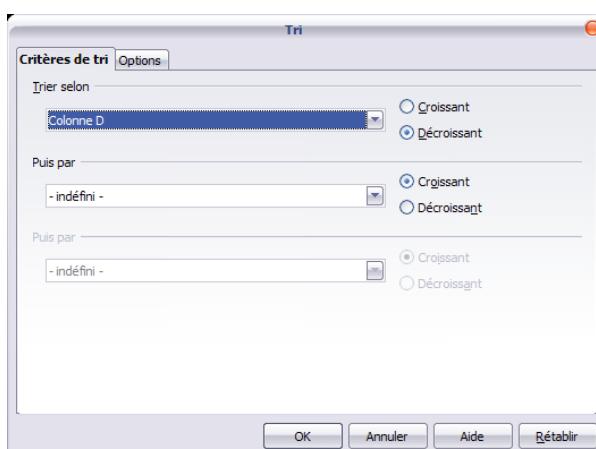
A	B	C	D	E	F	G	H
Id	Label	Corpus	Status	Browsed Pages Count	Flagged Pages Count	Node Type	TAG Dessin_amateur
6	ncs5	http://www.melakarnets.com	visite	3	0	site	non
7	ncs6	http://bruno.free.fr	visite	2	0	site	non
8	ncs7	http://www.pénélope-jolicoeur.com	visite	2	0	site	non
22	ncs22	http://www.bloglaurel.com	visite	7	0	site	non
23	ncs23	http://blog.chatbd.com	visite	2	0	site	non
25	ncs25	http://poopipanda.blogspot.com	visite	2	0	site	non
28	ncs28	http://www.evarollin.com	visite	2	0	site	non
31	ncs31	http://placeaman.canalblog.com	visite	2	0	site	non
34	ncs34	http://blog.wogah.com	visite	2	0	site	non
35	ncs35	http://figo.over-blog.com	visite	2	0	site	non
339	ncs338	http://missgally.com	visite	4	0	site	non
	592						
	593						

Trier d'après le nombre de pages visitées décroissant

Toujours dans le menu « Données » se trouve la fonction « Trier ». Sélectionnez toute la table et cliquez dans ce menu.



Une boîte de dialogue s'ouvre alors : choisissez selon quelle colonne vous voulez trier, et si vous souhaitez un ordre croissant ou décroissant.



Validez, le tri s'effectuera. Ici nous avons trié les sites par nombre de pages visitées décroissant. Vous pourriez tout aussi bien trier les URLs par ordre alphabétique pour les retrouver facilement.

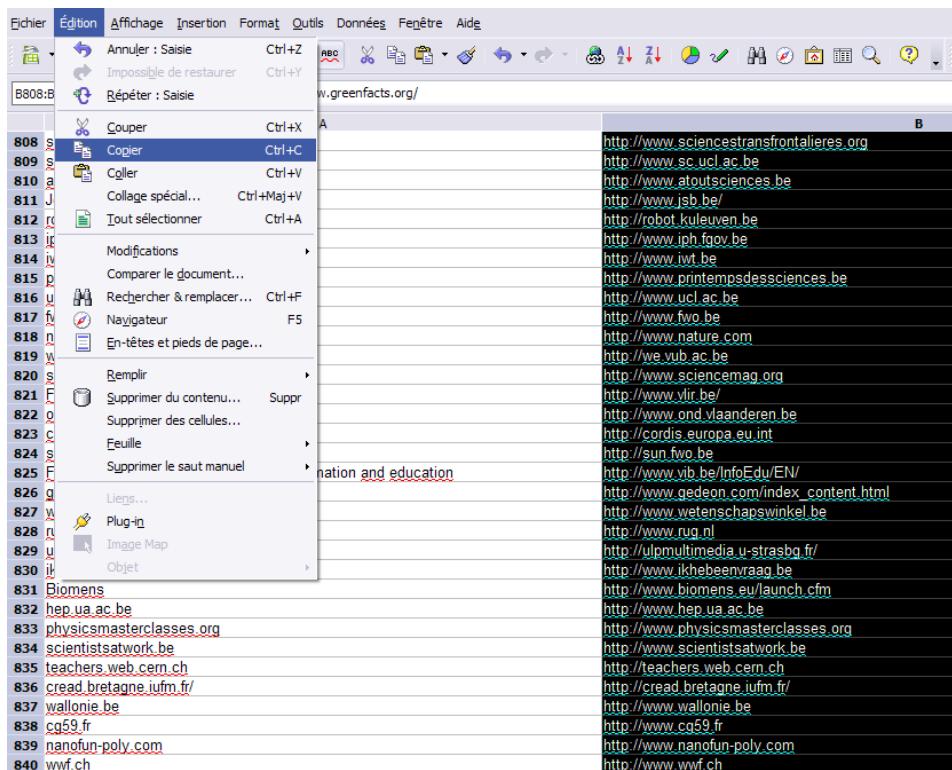
Le tableur vous sera donc utile pour sélectionner et trier les données, mais aussi pour enrichir vos données, les partager avec des personnes qui n'utilisent pas le

Navicrawler, pour produire des diagrammes et faire des analyses statistiques.

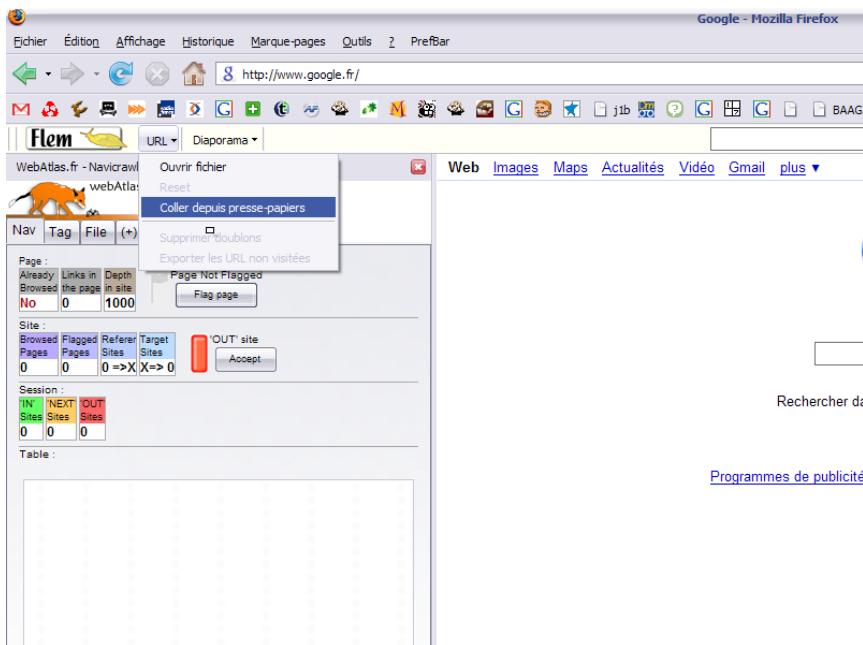
Les tableaux sont par ailleurs utilisés dans de nombreux cas pour stocker des listes d'URLs diverses. Tel organisme a par exemple répertorié ses partenaires dans un tableau disposant d'une colonne « site web ». Nous allons maintenant voir comment exploiter de telles listes d'URLs avec le Navicrawler. Pour ce faire nous allons utiliser une autre extension Firefox conjointement avec le Navicrawler : Flem.

Collecter une liste d'URLs avec le Navicrawler et Flem

D'abord, sélectionnez la liste d'URLs dans le tableur et copiez-la dans le presse-papier (ctrl+c ou menu « Edition » > copier).



Ensuite ouvrez le Firefox et activez le Navicrawler. Nous supposons que vous avez déjà installé Flem. Dans l'interface de Flem (une barre jaune au dessus de l'espace de navigation), dans le menu « URL », cliquez sur « coller depuis le presse-papiers ».



La liste d'adresses est désormais chargée dans Flem. Vous pouvez la visualiser en cliquant sur le menu déroulant principal de Flem.

	URL
0	http://www.sciencefronthoraires.org
1	http://www.sc.ugent.be
2	http://www.atousciences.be
3	http://www.jab.be/
4	http://robot.kuleuven.be
5	http://www.ipb.gov.be
6	http://www.wt.be
7	http://www.printempsdesciences.be
8	http://www.ud.ac.be
9	http://www.fivo.be
10	http://www.nature.com
11	http://vle.vub.ac.be
12	http://www.sciencemag.org
13	http://www.vri.be/
14	http://www.ond.vlaanderen.be
15	http://cordis.europa.eu/int
16	http://sun.fivo.be
17	http://www.vib.be/InfoEdut/EN/
18	http://www.gedeon.com/index_content.html
19	http://www.wetenschapswinkel.be
20	http://www.rug.nl
21	http://upmultimedia.u-strasbg.fr/
22	http://www.khebeenvraag.be
23	http://www.biomed.eu/launch.cfm
24	http://www.hep.u.ac.be
25	http://www.physicmasterclasses.org
26	http://www.scientistsatwork.be
27	http://teachers.web.cern.ch
28	http://cread.bretagne.ulfm.fr/
29	http://www.wallonie.be
30	http://www.cg59.be
31	http://www.nanofun-poly.com
32	http://www.wif.ch
33	http://www.antarctica.ac.uk
34	http://www.gipb.kva.se
35	http://www.unep.org
36	http://www.grida.no
37	http://nsidc.org
38	http://www.bbc.co.uk
39	http://www.iucn.org

A droite de ce menu déroulant, deux boutons vous permettent de naviguer dans la liste (« précédent » et « suivant »). Mais vous pouvez également balayer toute la liste à l'aide de la fonction « diaporama » de Flem, accessible dans le menu du même nom.

Balayez manuellement ou automatiquement la liste, et les pages visitées seront

naturellement mémorisées par le Navicrawler, puisqu'il est actif. Vous pouvez également balayer la liste d'URLs manuellement en effectuant un crawl en distance 0 et profondeur 1 à chaque fois, pour collecter plus de liens entre les URLs de la liste. Cette procédure est utile notamment pour savoir si les URLs de la liste sont reliées entre elles.

Exploiter les données dans un logiciel de graphes

Voir un graphe issu du Navicrawler avec Guess

La principale raison d'être du Navicrawler est de vous permettre de visualiser des explorations dans un logiciel de graphes. WebAtlas, l'association qui développe le Navicrawler, développe son propre logiciel dédié, Géphi, qui sera le principal outil de visualisation des données issues du Navicrawler. Malheureusement le développement de Géphi n'est pas terminé à l'heure actuelle, mais un autre logiciel existe néanmoins qui est tout-à-fait indiqué pour exploiter efficacement ces données : Guess. Nous allons donc donner des exemples de la façon dont on utilise un tel logiciel pour visualiser des graphes issus du Navicrawler.

NB : ce logiciel nécessite la plateforme Java pour fonctionner. Le téléchargement et les indications pour l'installer sont disponibles sur le site de ses concepteurs :

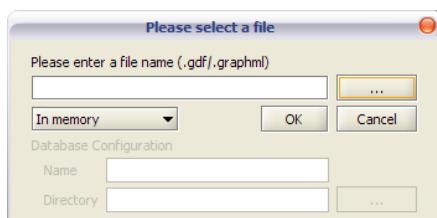
<http://graphexploration.cond.org/>

NB2 : si vous n'arrivez pas à lancer Guess, c'est que vous avez oublié d'éditer le fichier « guess.bat ». Il est nécessaire d'ouvrir ce fichier dans un éditeur de texte (comme le bloc-notes par exemple) et d'y inscrire à un certain endroit le chemin menant au répertoire dans lequel vous l'avez installé. La marche à suivre se trouve sur le site de Guess.

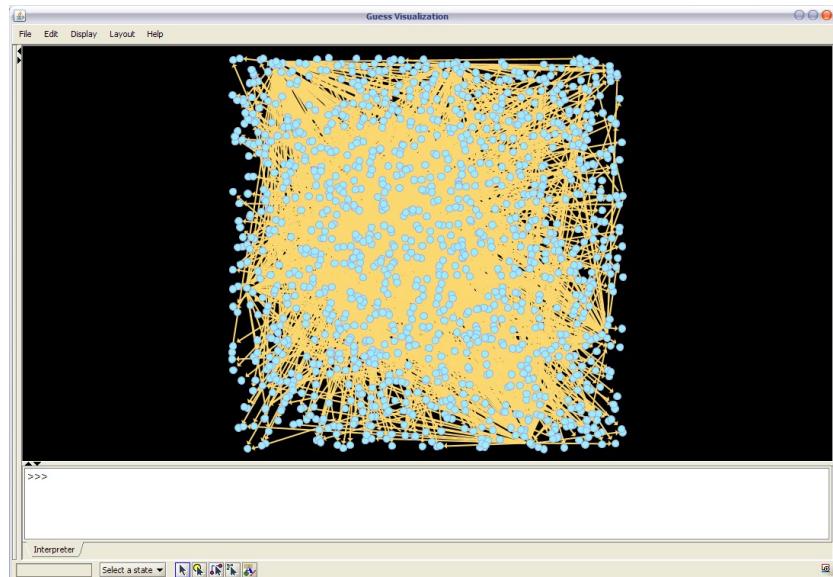
Voici donc comment visualiser vos données. Lancez Guess et dans la fenêtre d'accueil cliquez sur « Load GDF/GraphML ».



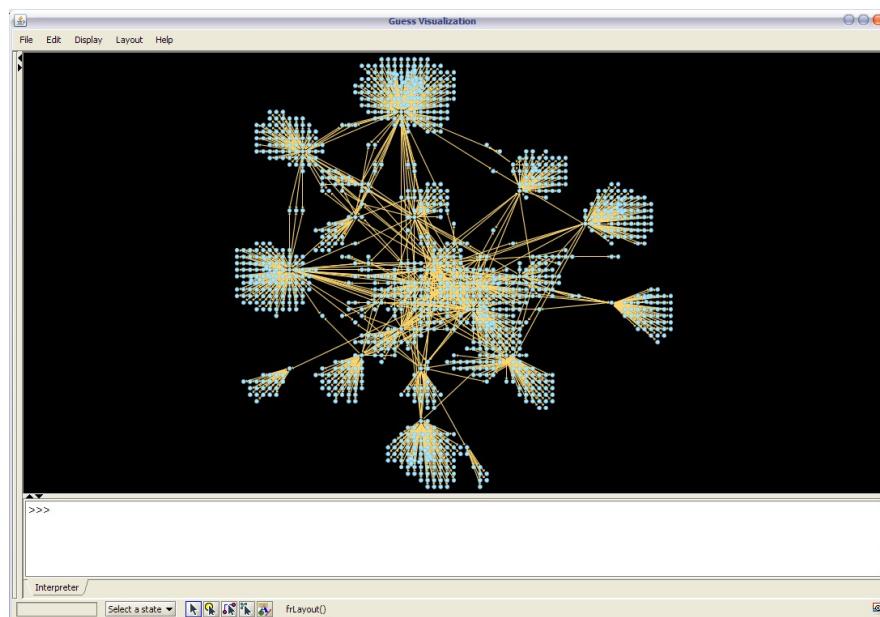
Dans le fenêtre de sélection du fichier à importer, cliquez sur les trois petits points et sélectionnez le fichier que vous souhaitez importer. **Ce fichier sera nécessairement un fichier GDF exporté par le Navicrawler.**



Le chemin du fichier s'affiche alors dans le champ textuel de cette fenêtre. Sans rien changer aux autres options, cliquez sur « OK ». Le temps que Guess se charge et importe le fichier, la fenêtre principale du logiciel s'ouvre et vous propose une première visualisation de vos données.

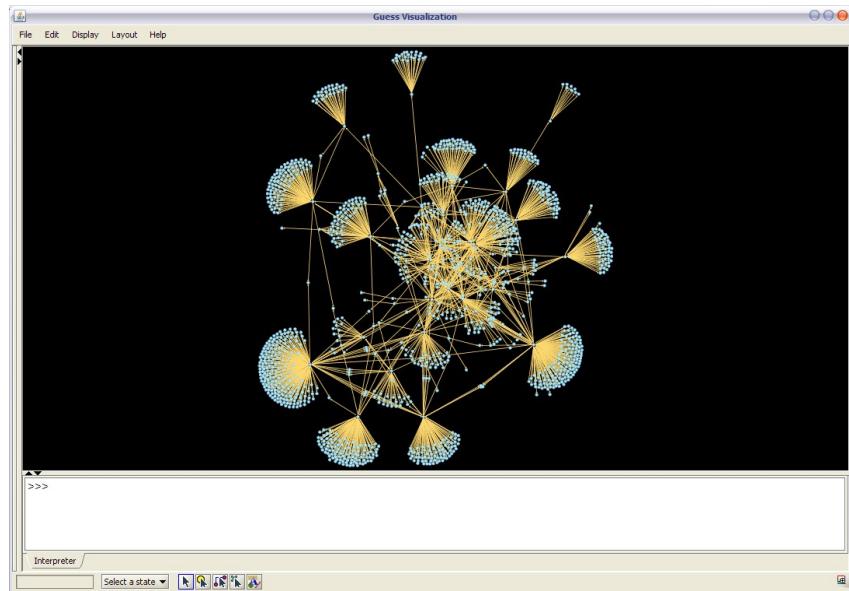


Le graphe apparaît comme un paquet carré et désordonné de noeuds reliés par des arcs. En effet, pour « voir » le graphe, il faut lui donner une forme. Les algorithmes qui donnent forme aux graphes sont appelés ici « Layouts » et fonctionnent pour la plupart sur le même principe : les noeuds se repoussent et les liens les retiennent comme des élastiques jusqu'à ce que le graphe se stabilise à un état d'équilibre. Effectuons une première spatialisation en cliquant dans le menu « Layout » et en cliquant sur « Fruchterman-Rheingold ». Après un temps d'attente plus ou moins long selon la taille du graphe, une nouvelle spatialisation apparaît.



Attention, la même spatialisation lancée plusieurs fois ne donne pas exactement le même résultat, bien que les formes générales se retrouvent la plupart du temps. De même, d'autres algorithmes donnent d'autres formes de

graphes. Essayons par exemple une autre spatialisation, « GEM » (toujours dans le menu « Layout »).



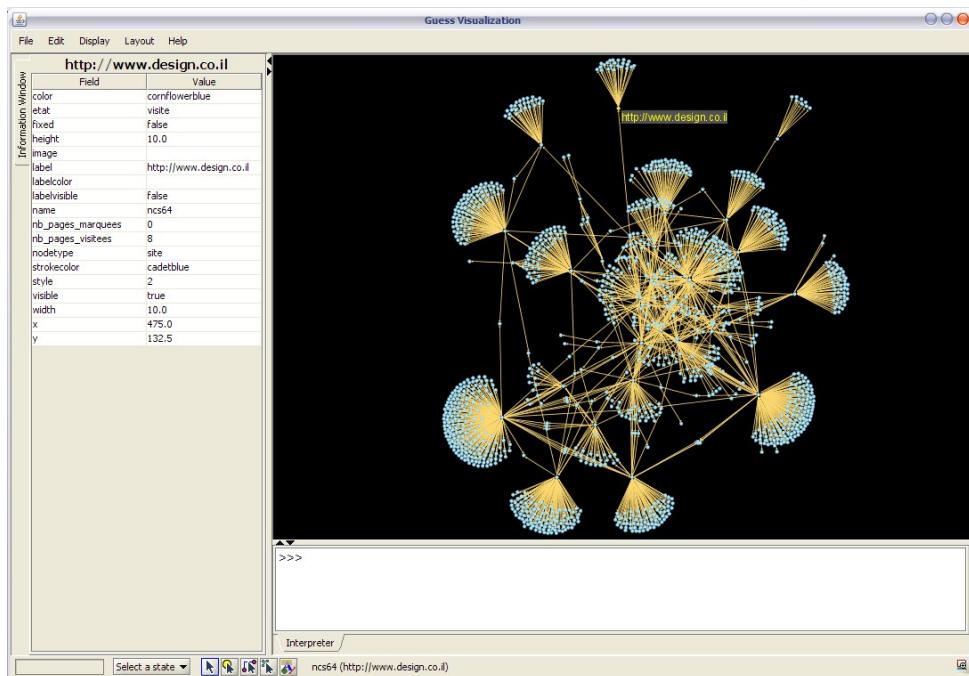
Un autre algorithme donne une autre forme au graphe. Cependant on remarquera que le centre du graphe reste le même, et que la morphologie globale de la périphérie aussi. Pourtant certaines excroissances qui étaient en haut peuvent se retrouver en bas etc. Nous vous encourageons à tester différentes visualisations pour vous faire une idée cohérente des propriétés « plastiques » de vos données. Les caractéristiques que l'on observe visuellement révèlent de propriétés structurelles du graphe, mais ces structures ne se donnent qu'à condition de manipuler suffisamment les données pour ne pas baser son interprétation sur les effets de bord de la spatialisation. La lecture de graphes nécessiterait un livre à part entière, nous n'entrerons donc pas dans ce sujet autant qu'il le faudrait. Toujours est-il que le principe de lecture fondamental repose sur la distinction entre cœur et périphérie, et sur l'observation fine de « qui est connecté à qui ». La dichotomie cœur/périphérie, mise en parallèle avec les modèles du web et la morphologie des domaines que nous avons évoqués plus haut, permet de hiérarchiser les données (selon leur connectivité et en extrapolant, leur généralité). L'observation fine quant à elle repose sur votre capacité à formuler une question précise à laquelle l'observation (et la manipulation !) du graphe vous permettra de répondre.

Pour l'instant le graphe n'est pas tellement lisible, nous allons donc effectuer quelques manipulations pour optimiser son apparence.

Modifier l'aspect visuel du graphe (introduction aux techniques)

Nous allons d'abord nous donner les moyens de manipuler les données accessibles dans Guess. Pour ce faire, cliquez sur le menu « Display » puis sur « Information window » pour afficher le panneau d'informations sur les liens et

les noeuds. Passez ensuite la souris sur un noeud pour afficher les informations correspondantes.



Dans le panneau d'informations, la colonne de gauche est commune à tous les noeuds et représente les différentes variables associées aux noeuds. La colonne de droite montre les valeurs prises par ces variables pour le noeud sur lequel on passe la souris. Ces données peuvent concerner les éléments graphiques (couleur, position...) mais aussi les données issues du Navicrawler. Ainsi sur l'exemple ci-dessus, le noeud actuellement survolé par la souris est de statut « incorporé » : la deuxième ligne nous informe que la variable « etat » du noeud a pour valeur « visite » (ancienne dénomination pour « incorporé »). De même on sait que ce noeud est un site (« nodetype »), qu'aucune page n'a été marquée et que 8 pages ont été visitées.

Connaître ces informations est essentiel puisque c'est sur elles qu'on se base pour intervenir sur le graphe. Ces interventions se font en lignes de commandes dans l'espace textuel situé en bas de la fenêtre. Les commandes simples fonctionnent sur le principe suivant :

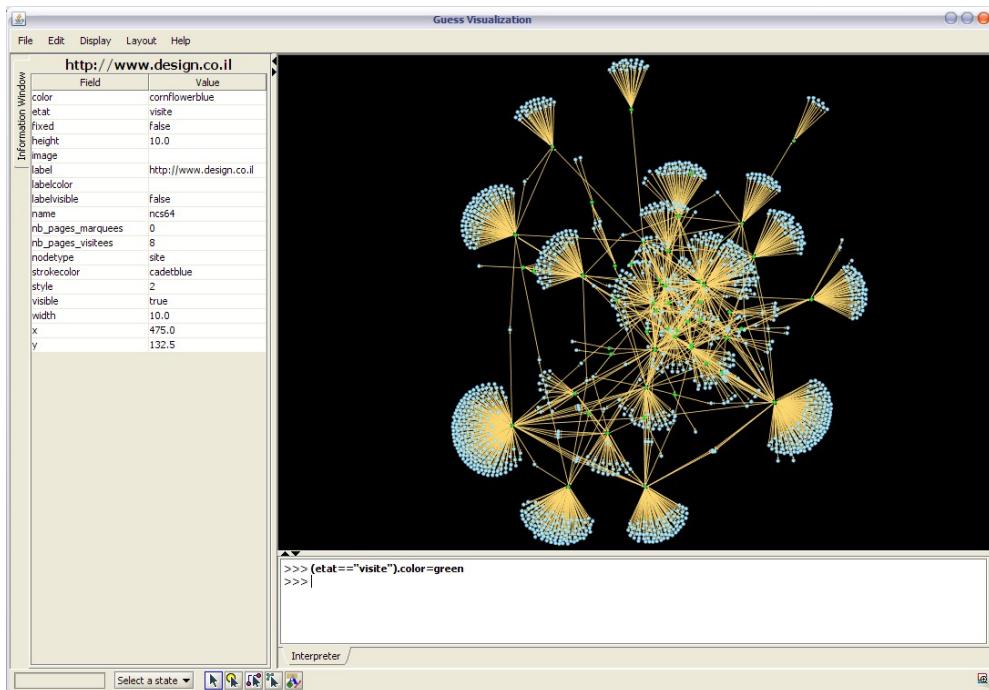
- D'abord on spécifie sur quels noeuds on veut agir en effectuant l'équivalent d'une requête. Le format est le suivant : (variable=="valeur") soit par exemple (etat=="visite") qui spécifie uniquement les noeuds incorporés.
- Ensuite on sélectionne une variable que l'on veut modifier, typiquement une variable graphique comme la couleur, et on affecte une valeur à cette variable.

Voici un exemple pour affecter la couleur verte aux sites incorporés :

```
(etat=="visite").color=green
```

Tapez cette ligne à la suite du prompteur « >>> » dans l'espace de lignes de commandes et appuyez sur entrée, si vous n'avez pas fait d'erreur de syntaxe la

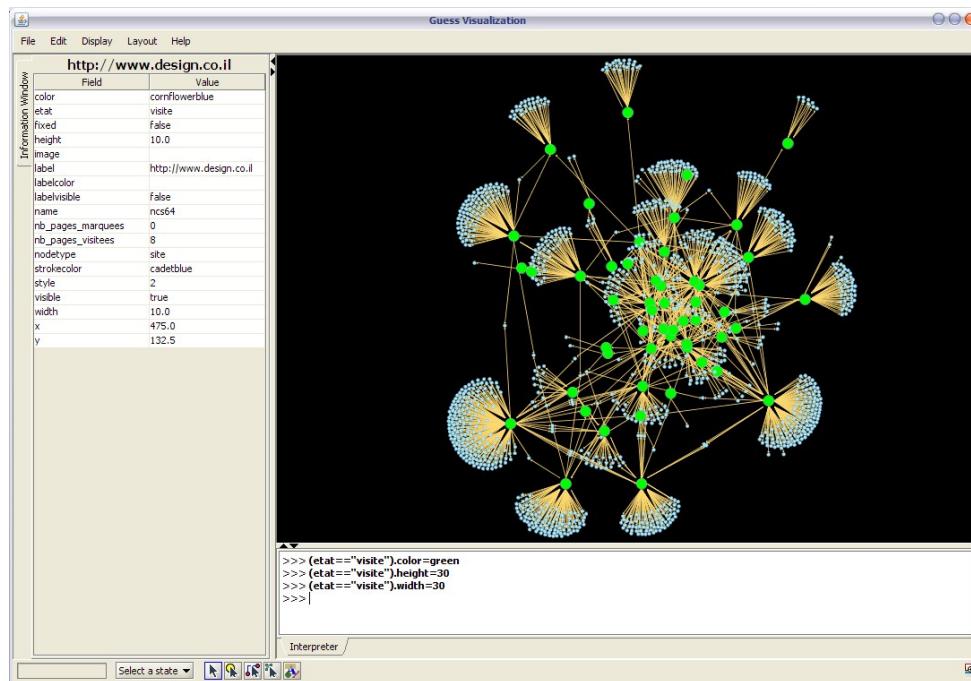
commande va s'exécuter.



Nous allons maintenant grossir les noeuds incorporés en agissant sur leurs paramètres « height » et « width » (hauteur et largeur).

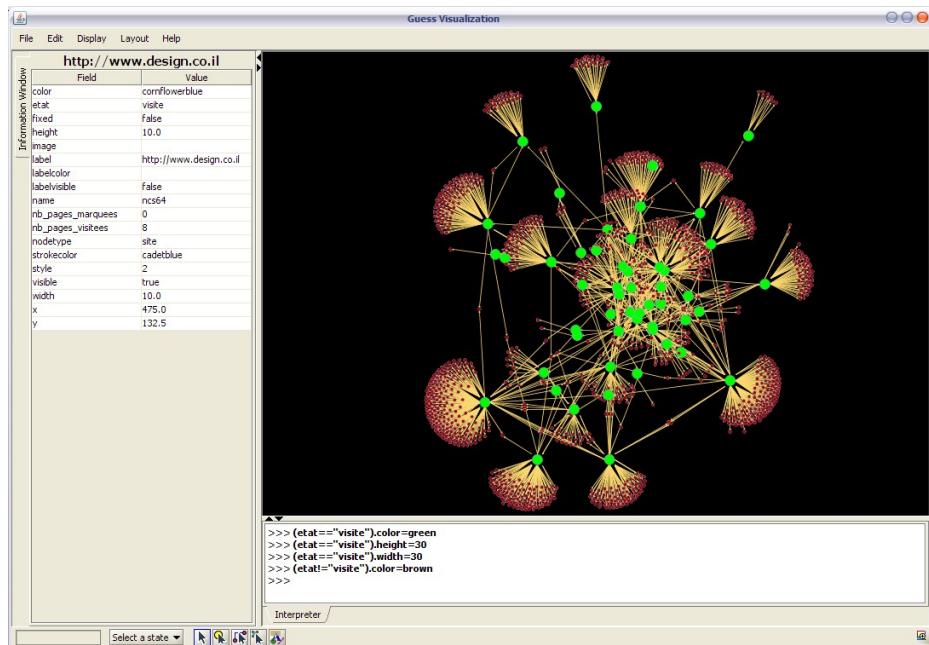
(etat=="visite").width=30

(etat=="visite").height=30



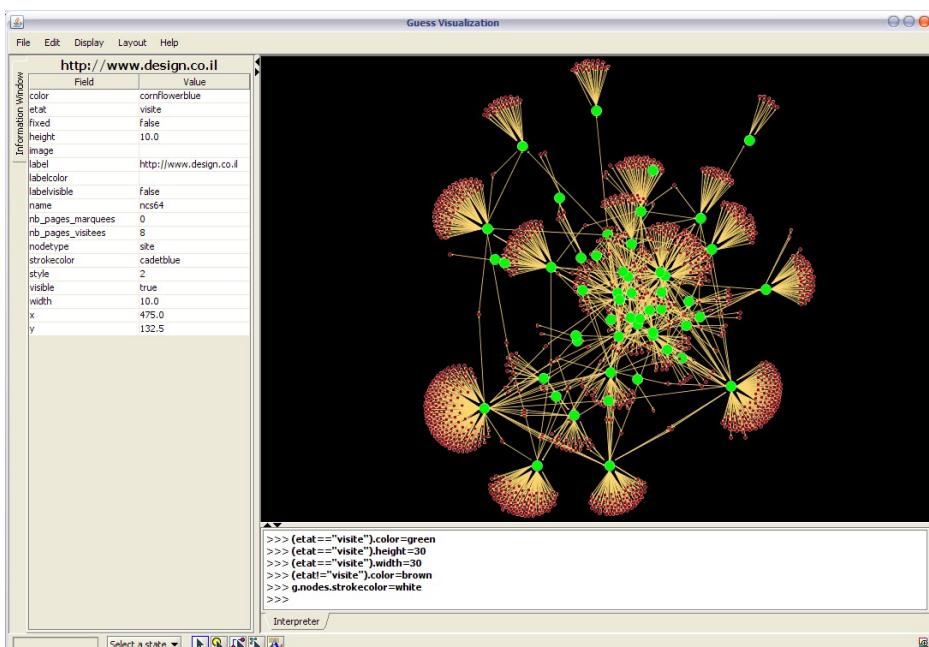
Nous allons maintenant colorer les autres noeuds en marron. Nous allons donc appliquer une couleur à tous les noeuds dont l'état est différent de « visite ». Le symbole « différent de » s'écrit : !=

(etat!="visite").color=brown



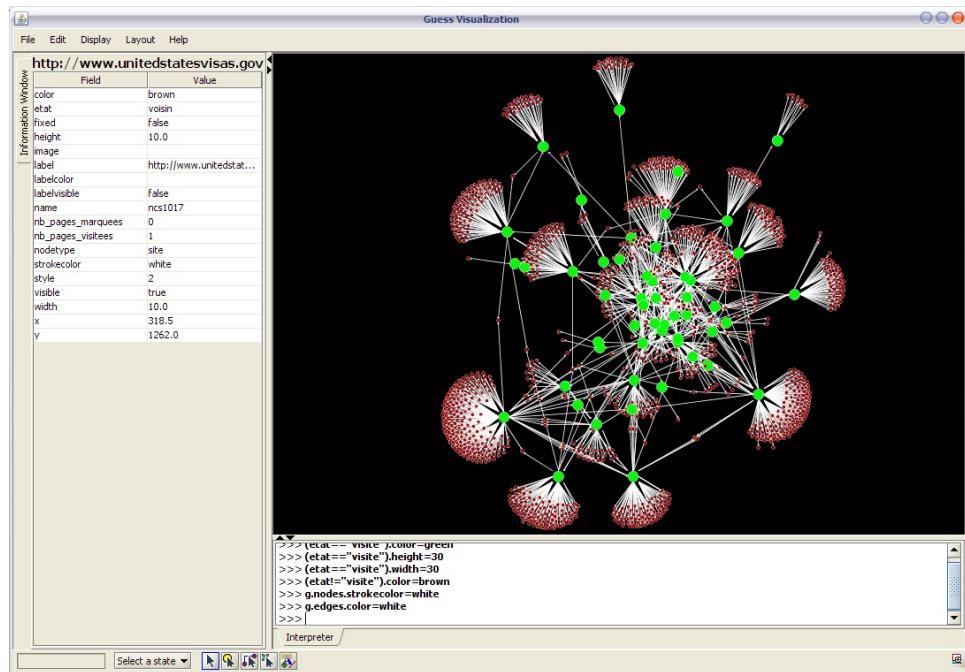
Nous allons maintenant agir sur tous les noeuds. « tous les noeuds » s'écrit « g.nodes ». Nous allons peindre en blanc le tour de tous les noeuds. La couleur du tour d'un noeud s'écrit « strokecolor » (vous pouvez la voir dans le panneau d'informations).

g.nodes.strokecolor=white



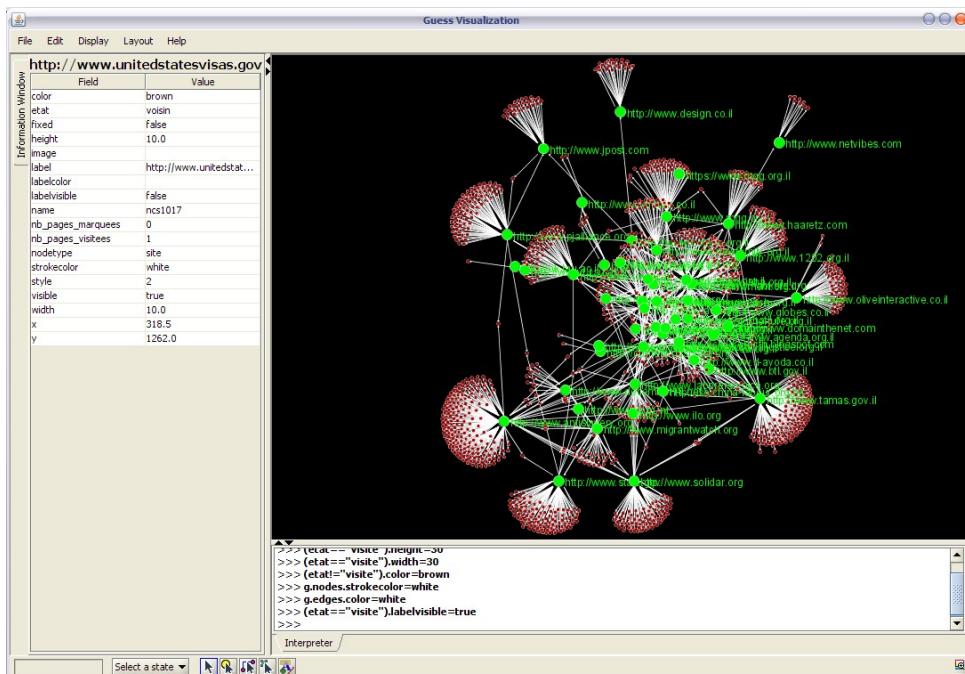
De même, nous allons peindre tous les liens en blanc avec cette commande que vous êtes maintenant capable de comprendre (« edge » signifie « arc ») :

g.edges.color=white



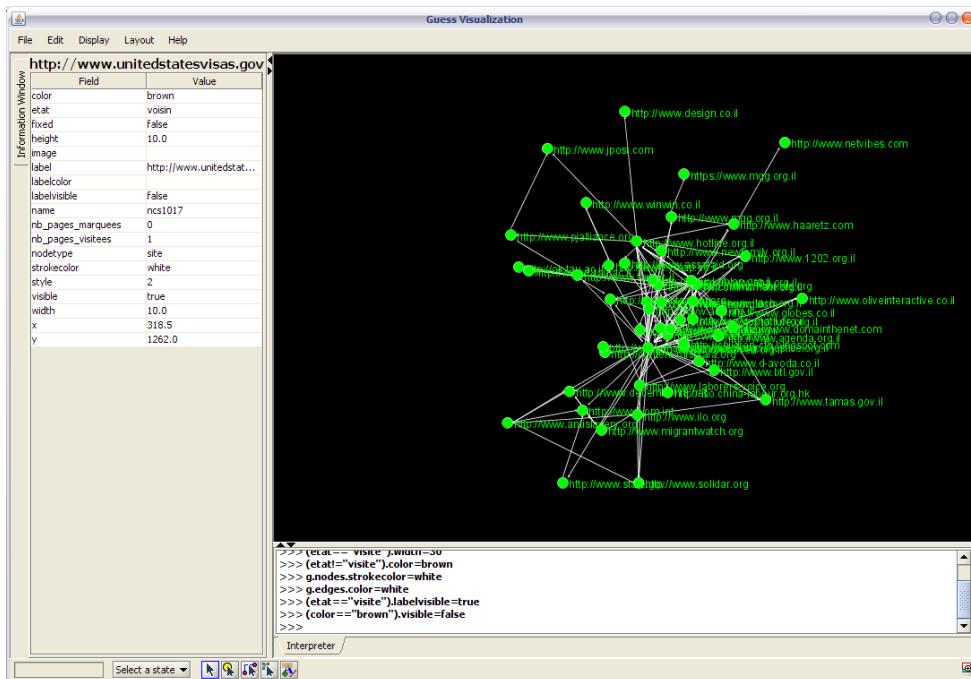
Maintenant nous allons afficher le nom des noeuds, mais uniquement pour les sites incorporés (car sinon il y en aurait trop).

(etat=="visite").labelvisible=true



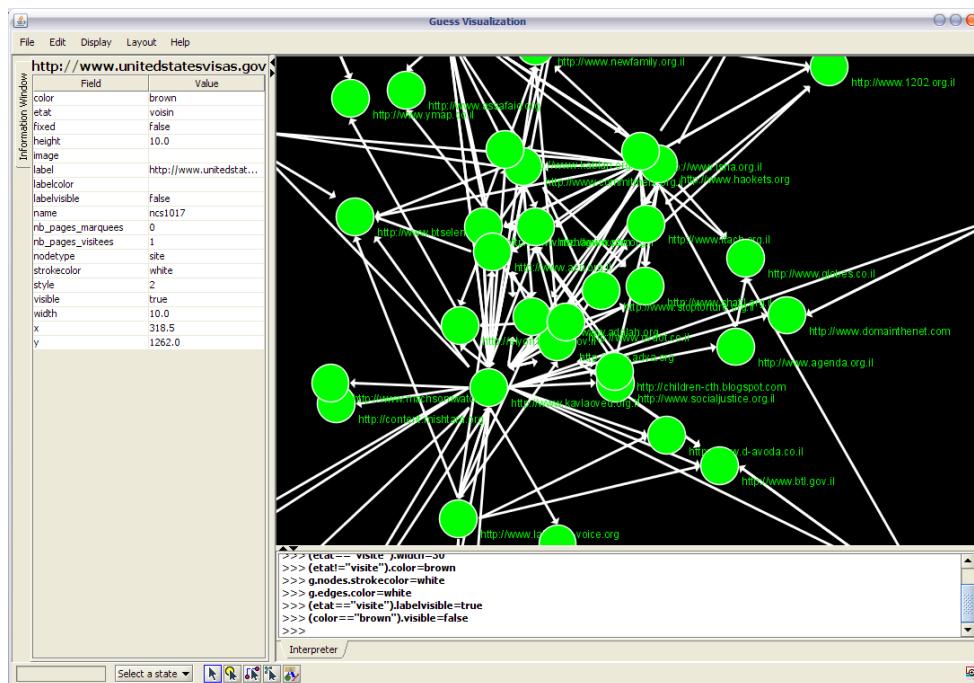
Les noeuds en marrons ne nous sont pas très utiles, nous allons donc les cacher. Si vous avez bien suivi cette série de manipulations, vous avez saisi que les noeuds en marron sont en fait les sites qui ne sont pas incorporés. Nous pouvons donc les sélectionner avec (`etat!="visite"`) mais cette fois-ci nous allons les sélectionner directement par la couleur :

```
(color=="brown").visible=false
```



On remarquera que des liens ont disparu : quand on masque un noeud, on masque en même temps tous les liens qui y sont connectés.

Nous allons maintenant zoomer sur le coeur du graphe pour voir de plus près. Le zoom s'active en effectuant un clic-droit et en glissant la souris vers la droite en gardant le bouton droit appuyé. De même le déplacement de la fenêtre s'effectue en glissant la souris avec le bouton gauche appuyé. Attention cependant, il ne faut pas cliquer là où il y a des noeuds ou des liens masqués, sinon ça ne fonctionne pas (c'est peu pratique mais c'est ainsi...).

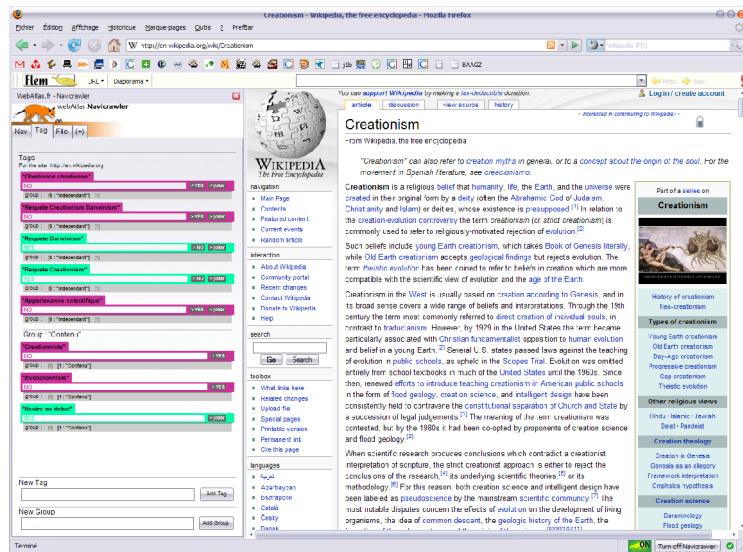


Les manipulations graphiques que vous serez amenés à effectuer avec Guess dépendent entièrement de ce que vous voulez montrer des données. Bien que les lignes de commande puissent sembler rebutantes dans un premier temps, elles sont un moyen efficace et polyvalent pour manipuler l'aspect graphique

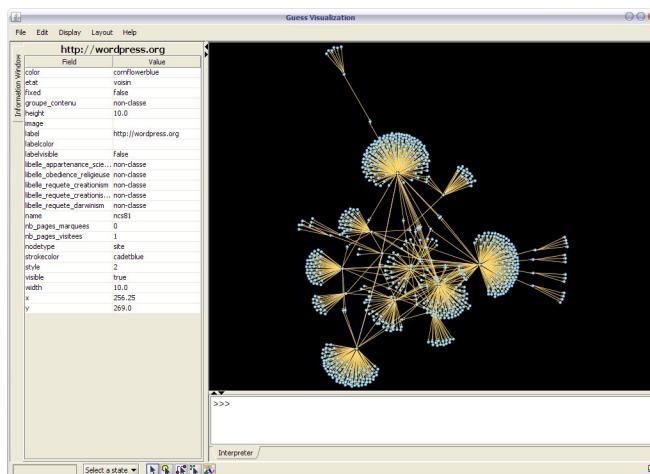
du graphe. Vous trouverez d'autres fonctions dans la documentation de Guess, disponible sur son site web. Les manipulations que nous avons effectuées ci-dessus n'ont pas une grande pertinence, elles sont à considérer comme un exercice qui vous aidera à effectuer vos propres manipulations et à comprendre la logique de fonctionnement de Guess. Nous allons maintenant voir comment exploiter plus intelligemment ces fonctions pour visualiser des libellés sur un graphe.

Visualiser les libellés

Nous allons prendre pour exemple l'exploration des 10 premiers résultats de Google pour trois requêtes, « creationism », « darwinism » et « creationisme darwinism ». Ces trois requêtes ont été collectées en même temps avec le Navicrawler, et nous avons ajouté des descripteurs sous la forme de libellés et de groupes de libellés.



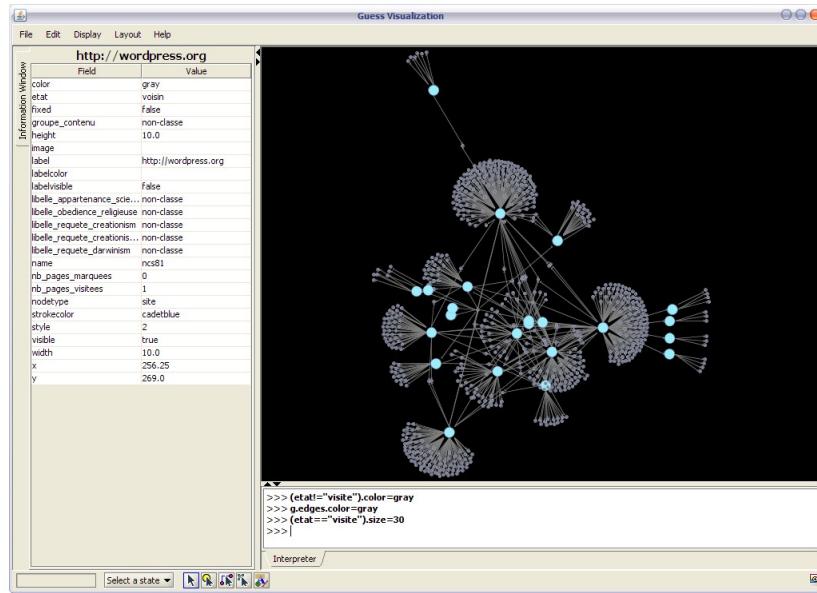
Ces informations vont maintenant nous permettre d'enrichir la visualisation pour tirer différentes conclusions. Le graphe est d'abord exporté avec le Navicrawler puis spatialisé avec l'algorithme GEM de Guess.



On effectue un rapide nettoyage visuel pour mettre en évidence les sites incorporés : passer les sites non-incorporés en gris, ainsi que les liens, et grossir

les sites incorporés.

```
(etat!="visite").color=gray  
g.edges.color=gray  
(etat=="visite").size=30
```



Lors de l'exploration des requêtes Google, nous avons utilisé un premier libellé pour noter si quel site était présent lors de la première requête, un deuxième libellé pour la deuxième requête et enfin un troisième libellé pour la troisième requête. Ces libellés sont indépendants, car un site peut être présent dans plusieurs requêtes. Nous allons visualiser ces données avec le code couleur suivant :

- Site présent pour « darwinism » : rouge
- Site présent pour « creationism » : vert
- Site présent pour la double requête « darwinism creationism » : bleu
- Les sites présents sur plusieurs requêtes sont marqués par les associations de couleurs :
 - « darwinism » et « creationism » : jaune
 - « creationism » et « darwinism creationism » : cyan
 - « darwinism creationism » et « darwinism » : magenta
 - Site présent dans les trois pages de résultats : blanc

Les trois premiers cas sont simples à mettre en place (attention à bien écrire le nom exact du paramètre, tel qu'il apparaît dans le panneau d'informations) :

```
(libelle_requete_creationism=="oui").color = green  
(libelle_requete_darwinism=="oui").color=red  
(libelle_requete_creationism_darwinism=="oui").color = blue
```

Les cas suivants sont plus complexes. Il faut passer par une boucle « for ». Nous n'expliquerons pas en détail cette syntaxe, car les exemples suivants devraient

vous permettre, en les recopiant, d'effectuer vos propres manipulations. Attention cependant, ne faites pas de fautes de frappes en tapant le nom des libellés (reportez-vous au panneau d'informations) et tapez bien les tabulations telles qu'elles apparaissent ici (une tabulation au premier saut de ligne, deux tabulations au deuxième saut de ligne).

for n in g.nodes:

```
if(n.libelle_requete_darwinism=="oui" and n.libelle_requete_creationism=="oui"):
    n.color=yellow;
```

<Passer trois fois à la ligne>

for n in g.nodes:

```
if(n.libelle_requete_creationism_darwinism=="oui" and n.libelle_requete_creationism=="oui"):
    n.color=cyan;
```

<Passer trois fois à la ligne>

for n in g.nodes:

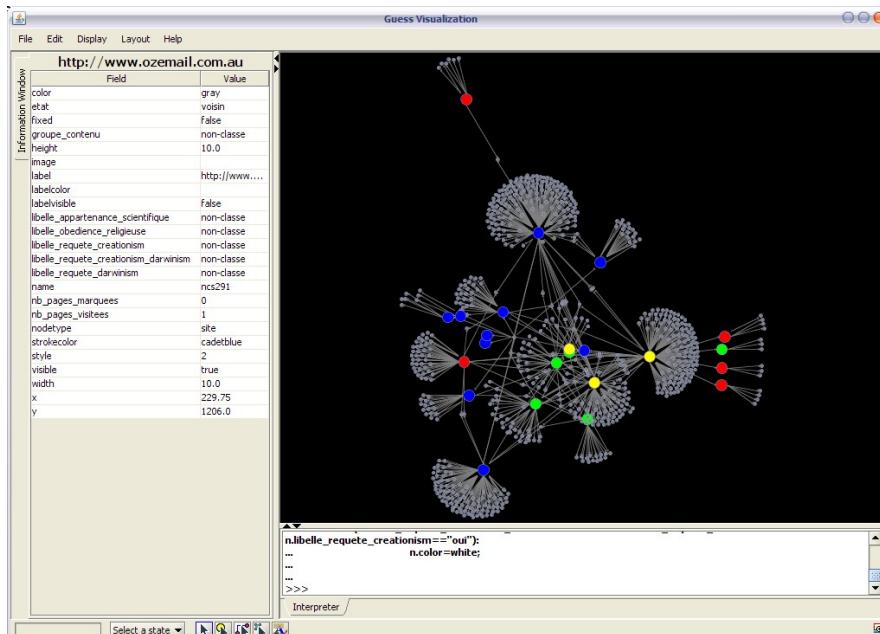
```
if(n.libelle_requete_darwinism=="oui" and n.libelle_requete_creationism_darwinism=="oui"):
    n.color=magenta;
```

<Passer trois fois à la ligne>

for n in g.nodes:

```
if(n.libelle_requete_darwinism=="oui" and n.libelle_requete_creationism_darwinism=="oui" and
n.libelle_requete_creationism=="oui"):
    n.color=white;
```

<Passer trois fois à la ligne>



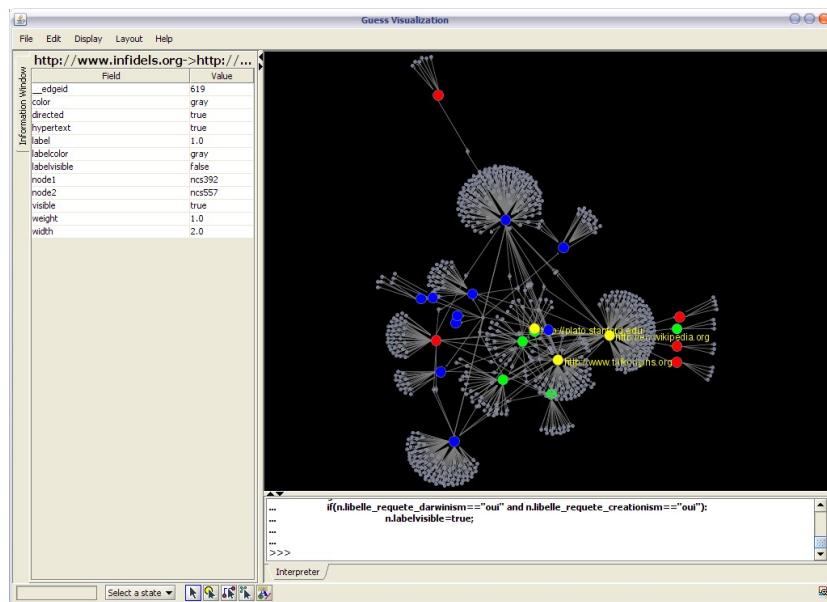
On observe d'ors et déjà l'absence de magentas, de cyans et de blancs. Les deux seules requêtes ayant des résultats communs sont donc « darwinism » et « creationism » (sites en jaune). Pourquoi ? Pour le savoir nous allons afficher

les libellés pour les sites concernés. De même, il faut passer par une boucle « for » :

for n in g.nodes:

```
if(n.libelle_requete_darwinism=="oui" and n.libelle_requete_creationism=="oui"):
    n.labelvisible=true;
```

<Passer trois fois à la ligne>

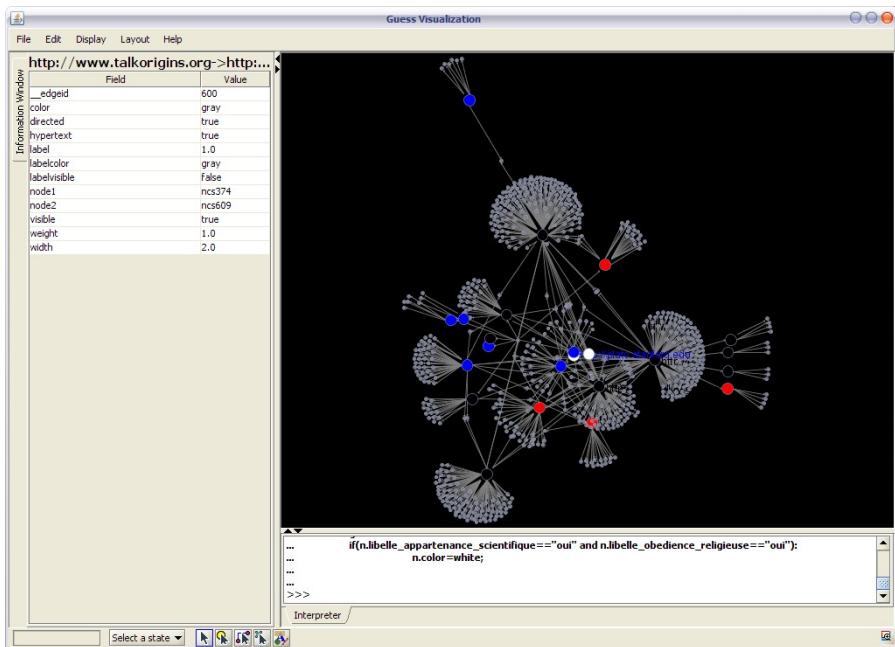


Parmi les trois sites on trouve Wikipedia : ce site très générique dispose d'une définition pour chacun des deux termes, mais n'étant pas dédié aux discussions ou controverses il n'est pas dans les ressources les plus pertinentes pour l'association des deux termes (l'interprétation pourrait être développée pour expliquer les deux autres cas).

On peut également remarquer que les sites bleus (requête « darwinism creationism ») sont relativement bien interconnectés, tandis que les rouges (« darwinism ») et les verts (« creationism ») ne le sont pas beaucoup hormis par les trois sites en commun.

Poursuivons notre investigation. Nous avons également utilisé deux libellés indépendants pour savoir si les sites sont d'une part d'obédience religieuse, d'autre part à teneur scientifique revendiquée. Nous allons utiliser un autre code couleur pour le visualiser : rouge=religieux, bleu=scientifique, blanc=les deux, noir=aucun. Voici le code exécutant ces manipulations :

```
(etat=="visite").color=black
(libelle_appartenance_scientifique=="oui").color=blue
(libelle_obedience_religieuse=="oui").color=red
for n in g.nodes:
    if(n.libelle_appartenance_scientifique=="oui" and n.libelle_obedience_religieuse=="oui"):
        n.color=white;
<Passer trois fois à la ligne>
```



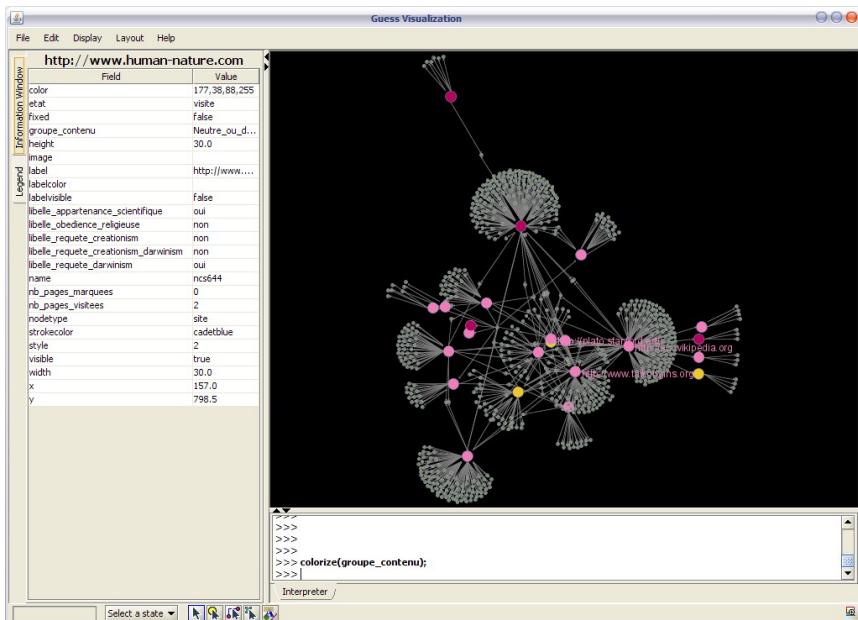
On retrouve au centre deux sites à la fois religieux et scientifiques (lesquels et pourquoi ?) et par ailleurs on trouve plus de sites scientifiques que de sites religieux. On note également un relativement grand nombre de sites ni religieux ni scientifiques (Quel est alors leur profil ?).

Enfin, nous avons utilisé un groupe de libellés pour situer le contenu du site : créationniste, évolutionniste, ou bien neutre/débat. Nous allons donc visualiser par des couleurs ce groupe de libellés. Pour ce faire nous allons utiliser une fonction dédiée de Guess. Voici le script :

```
colorize(groupe_contenu)
```

Cette ligne de commandes attribue une couleur aléatoire à chaque valeur prise par la variable « groupe_contenu ». Cette variable contient la valeur que nous avons choisie dans le Navicrawler pour chaque site. En regardant la fenêtre d'infos pour des noeuds de différentes couleurs, on retrouve la correspondance entre les couleurs et les libellés du groupe :

- Rose foncé = évolutionniste
- Rose clair = neutre ou débat
- Jaune = créationniste
- Gris = non-classé



On remarquera que sur les deux sites centraux scientifiques et religieux, l'un est neutre/débat, l'autre est créationniste. Par ailleurs on observe la prédominance des sites neutres ou de débat. (Nous n'entrerons pas plus loin dans l'interprétation ici).

Pour conclure cette partie dédiée aux libellés, résumons quelques points :

- Lorsque vous visualisez conjointement des libellés indépendants (ie. qui ne sont pas dans le même groupe), pensez bien à **toutes les combinaisons possibles**.
- Observez les **rapports quantitatifs** entre la distribution des différentes valeurs, mais observez aussi leur **connectivité**.
- Produisez différentes visualisations selon les libellés qu'il est pertinent d'associer entre eux, et **comparez-les** pour les interpréter.

Utiliser des scripts pour Guess⁶ (étudier la distribution des liens)

Dans Guess un certain nombre de fonctions peuvent être scriptées. Cela signifie que vous pouvez charger des fichiers capables de rajouter des fonctionnalités plus ou moins avancées à Guess. WebAtlas propose un tel fichier qui vous permettra d'effectuer diverses manipulations adaptées aux données issues du Navicrawler. Voici comment profiter de ce fichier pour analyser plus finement les données.

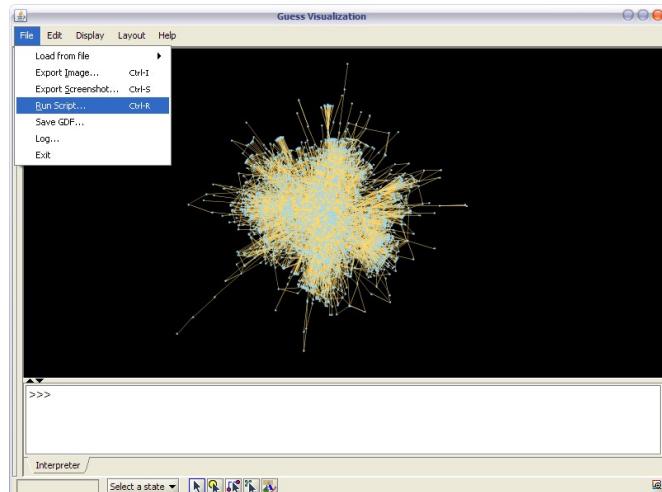
Ouvrez Guess et chargez un fichier GDF issu du Navicrawler, puis spatialisez-le. Par ailleurs, téléchargez le fichier de scripts à l'adresse suivante :

<http://www.webatlas.fr/ressources/download/webatlasScripts.py>

Maintenant il suffit de cliquer, dans le menu « File », sur « Run Script » et

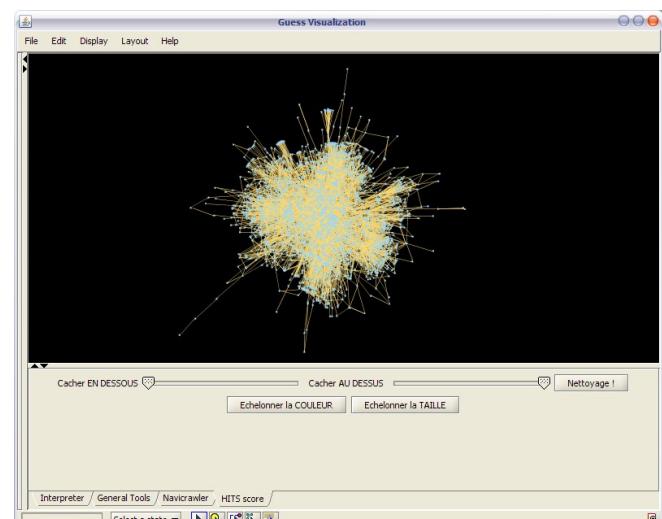
⁶ Page dédiée aux scripts WebAtlas pour Guess : http://www.web-mining.fr/blog/%5Buser%5D/scripts_pour_guess

d'ouvrir le fichier téléchargé pour charger les scripts dans Guess.

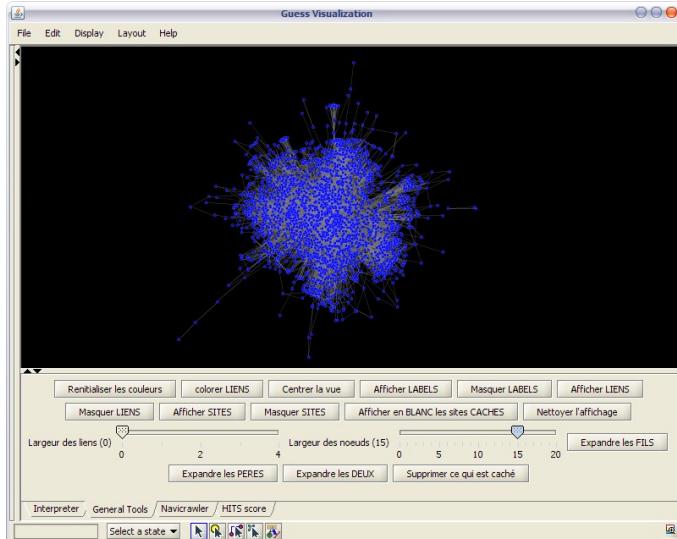


Une fois le script chargé, il se manifeste par la présence d'interfaces additionnelles dans la partie basse de la fenêtre. Ces interfaces sont réparties en trois nouveaux onglets qui concernent des lots de fonctionnalités différentes :

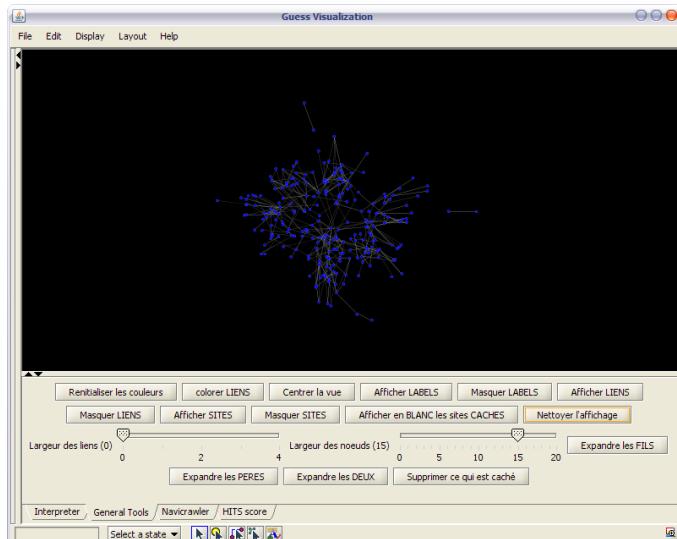
- General Tools, pour des manipulations génériques sur le graphe ;
- Navicrawler, pour des manipulations spécifiques aux données du Navicrawler ;
- HITS score, pour des manipulations relatives à la hiérarchie de la connectivité.



L'onglet « General Tools » permet d'effectuer rapidement diverses manipulations graphiques qui rendent le travail plus commode. Vous pouvez par exemple augmenter la taille des noeuds à 15, diminuer l'épaisseur des liens à 0, et réinitialiser les couleurs. Ces fonctions vous permettent notamment de **« remettre à zéro » les paramètres graphiques** du graphe pour faire une nouvelle colorisation.



Vous pouvez également « nettoyer » le graphe en cachant tous les noeuds qui ont zéro ou un seul lien, afin de **faire disparaître les structures dont les connexions sont peu significatives**. Vous remarquerez que lorsque vous nettoyez une fois, des noeuds disparaissent et donc cela diminue la connectivité d'autres noeuds, qui peuvent devenir à leur tour des noeuds à nettoyer. C'est pourquoi nous recommandons de cliquer plusieurs fois sur le bouton jusqu'à ce qu'il n'y ait plus de noeuds à nettoyer.

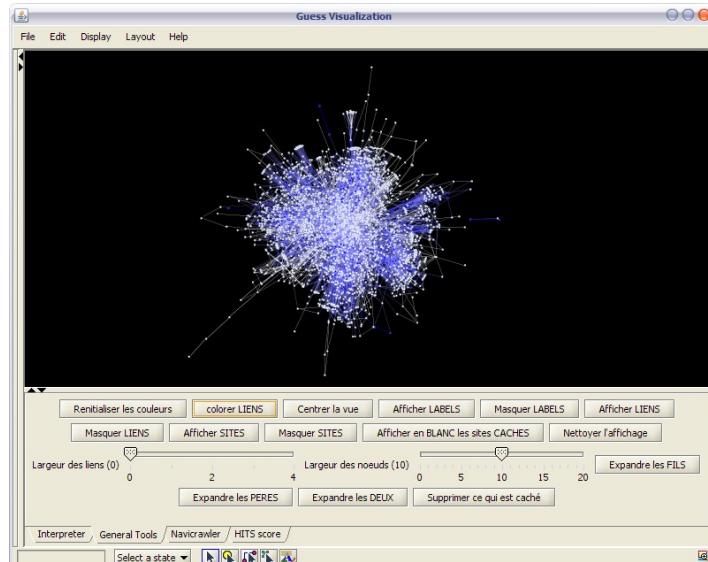


Nous verrons plus loin d'autres fonctions qui n'affichent qu'une partie du graphe. Lorsque c'est le cas, l'onglet « General Tools » permet d'effectuer différents traitements pour mettre en valeur ce sous-graphe :

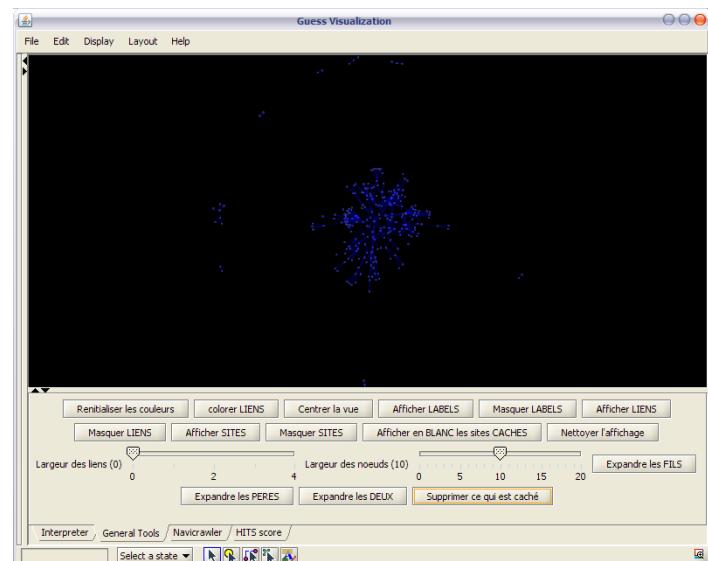
- Vous pouvez jouer sur les couleurs pour faire ressortir le sous-graphe
- Vous pouvez augmenter le sous-graphe visible en ajoutant les noeuds pointés, en ajoutant les noeuds qui pointent vers lui, ou les deux à la fois
- Vous pouvez supprimer définitivement les noeuds cachés et refaire une spatialisation

Pour jouer sur les couleurs, il faut d'abord cliquer sur « Afficher en BLANC les sites CACHES ». Afin d'améliorer le rendu des couleurs, il est aussi utile de

cliquer sur « colorer LIENS » pour affecter à chaque lien la couleur moyenne entre les deux couleurs des noeuds aux deux bouts de ce lien (c'est visuellement plus homogène). Des « structures fortes » apparaissent donc en couleur au milieu du blanc des sites peu connectés.



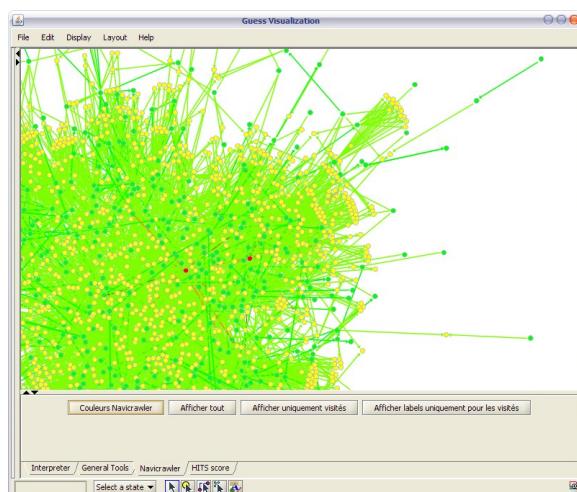
Voyons maintenant comment ne conserver que les structures fortes pour les observer indépendamment du reste ; pour ce faire revenons à un graphe nettoyé. Cliquez sur « Supprimer ce qui est caché » et effectuez une nouvelle spatialisation : les noeuds nettoyés sont complètement éliminés et ne jouent plus dans la spatialisation. Attention cependant, pour revenir en arrière et récupérer les noeuds effacés, vous serez obligés de recharger les données initiales.



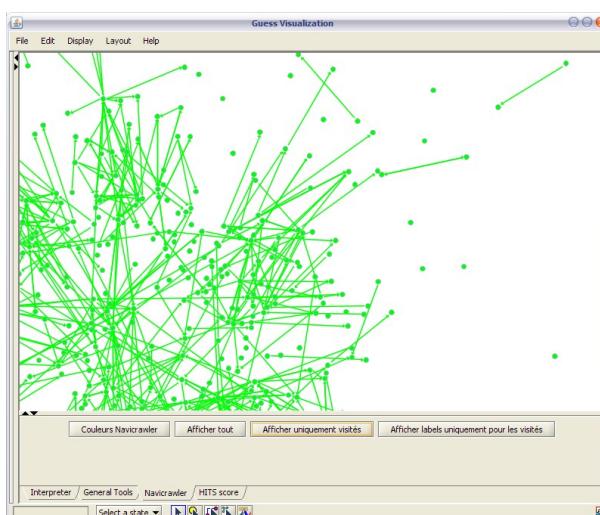
Ce que nous appelons ici les « structures fortes » que nous révèlent des manipulations sur le graphe, correspondent aux zones du graphe dans lesquelles les parcours sont « emmêlés ». Autrement dit, puisqu'on a supprimé les « culs-de-sac », il ne reste que les noeuds qui sont pris dans des réseaux relativement complexes. Les limites de cette technique résident dans le fait qu'un petit graphe n'aura presque pas de structures fortes tandis qu'un grand

graphe aura de telles structures en trop grandes quantités ; l'onglet « HITS score » permet une meilleur hiérarchisation des données. Par contre il est intéressant d'utiliser cette fonction soit pour **réduire des graphes** trop grands, soit pour **déetecter des « coeurs locaux »** c'est-à-dire des zones plus denses qui correspondent aux centres des sous-domaines, soit pour repérer des **grumeaux de ressources** qui forment des petites familles indépendantes dans le tissu du web (par exemple quand un webmestre a décidé de créer plusieurs sites inter-reliés au lieu des habituelles sections d'un même site). En effet, cette technique ne conserve que les noeuds qui forment des groupes. Ainsi deux noeuds qui se pointent mutuellement seront conservés (c'est le groupe minimal) tandis qu'une chaîne de sites qui pointent chacun le suivant sera éliminée, de même les sites prochains pointés par un seul site incorporé sont systématiquement nettoyés car ils ne constituent pas un groupe. Lorsque des paquets étendus mais disjoints apparaissent, ce sont des coeurs locaux, des zones appartenant à des domaines différents, ou le plus souvent de simples grumeaux de ressources.

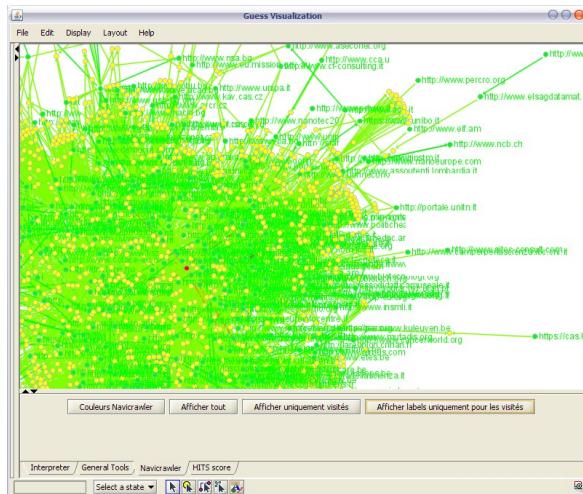
Voyons maintenant l'onglet « Navicrawler ». Son premier bouton permet de colorer le graphe selon le code-couleur habituel : vert pour les sites incorporés, orange pour les sites prochains, et rouge pour les sites écartés.



Il permet également de n'afficher que les sites visités, utile lorsque vous n'avez pas fait la sélection lors de l'export depuis le Navicrawler.

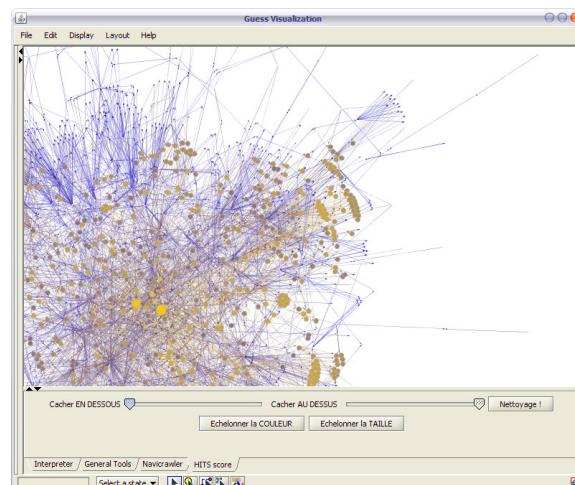


La dernière fonction, elle aussi très simple, vous permet de n'afficher le nom que pour les sites visités... C'est une fonction à exploiter lorsque le graphe est trop dense pour afficher tous les noms.



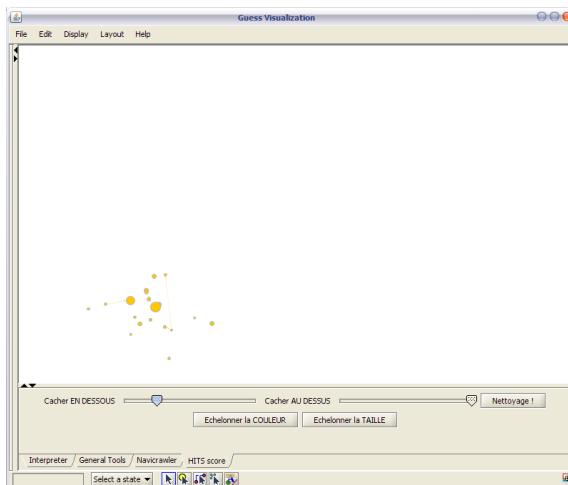
Enfin, le troisième onglet « HITS score » est dédié à la hiérarchie de la connectivité. Le hiérarchie la plus simple repose sur le fait de compter les liens. L'algorithme HITS prend en compte le fait qu'un lien avec un noeud très connecté « compte plus » qu'un lien avec un noeud peu connecté (renforcement mutuel). L'algorithme originel différencie les liens entrants et les liens sortants, mais l'implémentation dans Guess ne prend pas en compte la direction des liens. Plus le score de HITS d'un noeud est élevé, plus le noeud est important dans la hiérarchie de la connectivité. Autrement dit, **les grands scores de HITS sont les noeuds du cœur ou de la couche haute**.

Deux façons complémentaires de manifester l'importance des noeuds dans la hiérarchie de la connectivité sont l'application d'un dégradé de couleurs et l'affectation de tailles différentes. Cliquez sur « échelonner la couleur » puis « colorer les liens » (dans l'onglet General Tools) et les noeuds les moins importants (dans la hiérarchie de la connectivité) apparaîtront dans des tons bleus, tandis que les plus importants apparaîtront dans des tons jaunes. Cliquez sur « échelonner la taille » et les plus importants apparaîtront en plus gros.



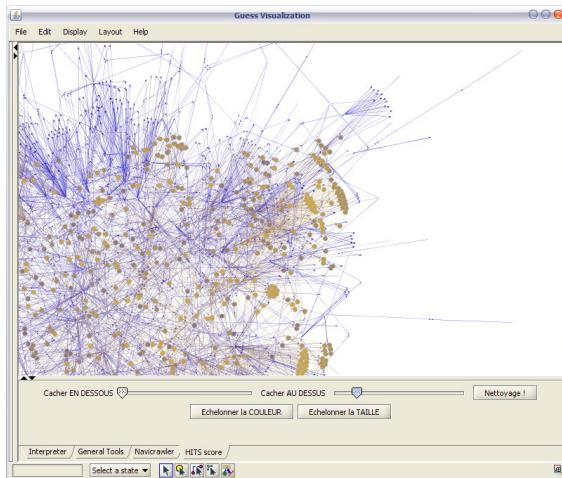
Les deux curseurs fonctionnent conjointement et vous permettent de n'afficher

que certains noeuds selon leur connectivité. Glissez le curseur de gauche vers la droite et les noeuds avec des petits scores seront masqués. Vous remarquerez que les noeuds visibles sont jaunes et de plus grande taille, puisqu'ils ont des scores importants. Selon l'étroitesse de votre sélection, les noeuds visibles sont le coeur et la couche haute du domaine, voire une partie de la nébuleuse. Le résultat classique attendu est que ces noeuds sont assez largement connectés ; si ce n'est pas le cas c'est probablement que le domaine traité n'est pas une communauté constituée : les acteurs du domaine ne peuvent pas avoir une conscience commune du territoire qu'ils occupent si les noeuds les plus connectés (autorités, leaders et relais d'opinion...) ne se connaissent et ne se reconnaissent pas mutuellement (en se liant hypertextuellement). **Si les noeuds les plus connectés du domaine ne sont pas liés entre eux, l'ensemble des acteurs ne se vit pas comme une communauté.**

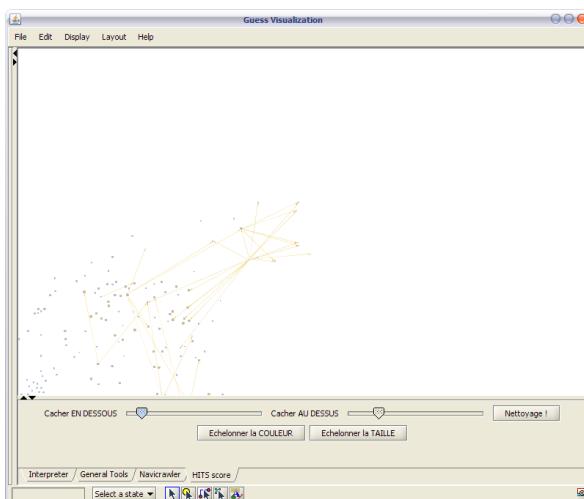


En remettant le curseur de gauche tout à gauche, vous affichez de nouveau l'intégralité du graphe. Glissez alors le curseur de droite sur la gauche : en faisant ainsi vous masquez maintenant les noeuds les plus connectés (la couche haute et le coeur, voire plus selon le dosage...). Les noeuds qui restent sont donc les moins connectés. Ceci permet de répondre à une question intéressante : les « petits » sites, ceux qui n'ont pas beaucoup de liens, sont-ils connectés uniquement aux « gros » sites ou bien sont-ils aussi connectés entre eux ? **Si les sites les moins connectés sont néanmoins connectés entre eux, alors il y a une certaine animation communautaire dans le domaine** (ce peut être le cas partout ou bien dans certaines zones seulement...). En effet il est courant que les « petits » sites soient principalement connectés aux « gros » sites ; le phénomène est assez simple à comprendre si l'on se dit que les sites d'acteurs qui n'ont pas beaucoup d'activités ou de renommée, ou sur les sites qui n'ont pas beaucoup de pages pour diverses raisons, les liens vont surtout vers des sites de forte notoriété, d'une part parce que plus un site est connu plus c'est facile de le découvrir et de mettre un lien vers lui, et d'autre part parce que cet attachement ramène une forme d'institutionnalisation (à l'échelle du web) dans le « petit » site qui en a besoin. Le fait que les petits sites se lient souvent aux gros, qui peut en partie

s'expliquer par le fait que les liens jouent le rôle de référence (au sens d'une référence dans un CV), est parfois désigné sous le nom d'attachement préférentiel.



En utilisant les deux curseurs, vous pouvez également faire apparaître certaines « couches de connectivité », c'est-à-dire les noeuds dont le score de HITS est compris entre deux bornes, fixées par les curseurs gauche et droite.



Ces manipulations sur le graphe vous permettent de comprendre la hiérarchie de la connectivité, mais aussi sa distribution. Nous allons maintenant vous donner quelques repères pour fonder votre interprétation.

Interpréter la distribution des liens

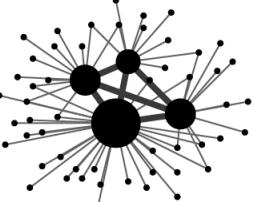
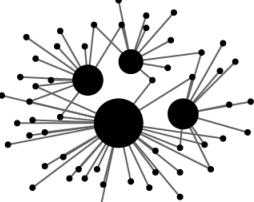
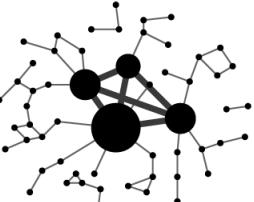
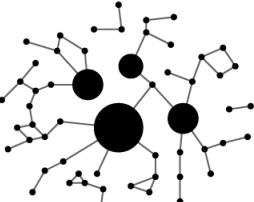
La théorie des graphes prévoit que pour un « réseau à invariance d'échelle » la distribution du nombre de noeuds suit une « loi de puissance » : sans rentrer dans les détails, ceci signifie qu'il y a toujours une petite minorité de noeuds très fortement connectés et une large majorité de noeuds peu connectés. Le web est globalement un tel réseau, et la plupart du temps les sous-graphes du web le sont aussi. Autrement dit il est **très probable** (voire presque certain) que **vos données laisseront apparaître une zone très dense**. Si ce n'est pas le cas alors il sera indispensable d'interpréter ce fait car il est fortement inhabituel (par exemple, votre domaine pourrait être l'une des rares

communautés complètement distribuées, autrement dit « égalitaires », du web) ; cependant nous vous recommandons d'envisager la possibilité d'un problème technique ou méthodologique (par exemple vous avez très peu de sites incorporés et les sites prochains biaisen le corpus)...

A supposer donc que vos données présentent une zone dense, il vous faudra examiner les points suivants (comme expliqué ci-dessus) :

- Les sites de la zone *dense* sont-ils connectés entre eux ?
- Les sites des zones *peu denses* sont-ils connectés entre eux ?
- Les sites de la zone *peu dense* sont-ils connectés avec la zone *dense* ?

La typologie ci-après vous permettra de vous situer parmi les différents cas. Deux remarques cependant. Tout d'abord les deux faits inattendus (et donc essentiels pour l'interprétation s'ils se présentent) sont d'une part le fait que les site peu connectés soient connectés entre eux, d'autre part le fait que les sites très connectés ne soient pas connectés entre eux. Or attention aux effets de masque pour tester ces deux possibilités : étant donné que les sites très connectés sont systématiquement connectés avec les sites peu connectés, il faut éliminer ces connexions pour faire apparaître l'un ou l'autre de ces deux cas. C'est ce pourquoi les curseurs de l'onglet « HITS score » sont pertinents : sans eux, c'est-à-dire simplement en regardant le graphe, vous ne verrez pas ces éléments de structure. Par ailleurs, gardez à l'esprit que la typologie présente des cas idéaux et que la plupart des sets de données sont un mélange qui se rapproche plus ou moins d'un type ou d'un autre ; de même les interprétations proposées sont des suggestions d'hypothèses que vous devrez vérifier par le contenu des sites. Pour cette raison les éléments de structure que vous aurez à interpréter seront souvent les suivants : **pourquoi ces « petits » sites *sont-ils connectés entre eux* ?** (et pas ceux-là) **pourquoi ces « grands » sites *ne sont-ils pas connectés entre eux* ?**

	<p> Zone dense connectée (les sites très connectés ont beaucoup de liens entre eux) Cette zone est le cœur, c'est un indice d'une centralité communautaire.</p>	<p>Zone dense NON connectée (les sites très connectés ne sont connectés qu'aux sites peu connectés) Cette zone est le (ou les) centre(s) : lieu de notoriété "autocratique".</p>
 Zones peu denses en étoile (Les sites peu connectés sont liés seulement aux sites très connectés) Domaine très hiérarchique, peu communautaire.	1  <p>Communauté hiérarchique : des sites existent autour d'un cœur constitué mais l'activité communautaire dépend des acteurs majeurs.</p>	2  <p>Domaine disparate : on observe bien des centres mais les acteurs du domaine ne communiquent pas entre eux.</p>
 Zones peu denses en maillage* (Les sites peu connectés sont liés entre eux) Domaine doté d'une bonne animation communautaire, reconnaissance mutuelle.	3*  <p>Communauté active : la reconnaissance mutuelle est présente à tous les niveaux, il y a activité et conscience communautaire.</p>	4*  <p>Communauté horizontale : le ou les centres ne communiquent pas, pourtant une activité communautaire existe entre les acteurs.</p>

* /!\ Attention : contrairement à l'illustration, le maillage co-existe souvent avec la distribution en étoile. Il faut donc éliminer les liens en étoiles pour savoir s'il y a ou non un maillage, indice d'activité communautaire.

Ces liens entre sites très connectés et sites peu connectés se prêtent relativement bien à l'interprétation en termes de liens forts et de liens faibles (*à la façon de Granovetter*). Dans cette perspective les liens forts sont les liens avec les sites les plus connectés, et les liens faibles sont les liens entre les sites les moins connectés. Ceci en toute cohérence avec le fait que nous proposons d'interpréter les liens entre « petits » sites comme l'indice d'une activité communautaire (présence de liens faibles). Mais attention, cette interprétation ne tient pas debout si elle n'est pas corroborée par des indices tirés du terrain lui-même ; et ceci est une remarque générale : les logiciels de graphes sont très puissants pour interpréter des structures qui autrement nous resteraient cachées, mais ils ne valent que par les hypothèses qu'ils permettent de formuler et qu'il faut éprouver par ailleurs. Ainsi **vous devrez revenir au terrain** (les sites web qui figurent dans le graphe) **pour valider les résultats de l'analyse des graphes**. Une propriété fondamentale de ce type de données est la possibilité qui vous est offerte d'effectuer des allers-retours entre grandes quantités de données (quantitatif) et analyse fine des contenus (qualitatif) et ce afin de vous permettre de prendre en compte le tout comme contexte du particulier et le particulier comme élément articulé du tout. L'exploitation de cette possibilité est la condition *sine qua non* du succès de ce type d'analyse, que nous appelons pour cette raison (en reprenant les termes de Bruno Latour) études quali-quantitatives.

Scénarios d'usage

Dans cette partie nous allons aborder par l'exemple différentes méthodes qu'il est possible de mettre en place avec le Navicrawler. Sous la forme de scénarios d'usage, nous allons lister les différentes tâches nécessaires et leur enchaînement pour plusieurs types de terrains, jusqu'à quelques éléments d'interprétation.

Le projet privilégié du Navicrawler, mais aussi le plus difficile, est l'étude d'un domaine complet du web. Nous allons commencer par détailler le scénario le plus classique, à savoir l'exploration sans objectif particulier, puis nous allons aborder pas à pas les scénarios qui visent à répondre à des objectifs précis, jusqu'à l'étude complexe d'un domaine.

Le premier scénario nous servira de « colonne vertébrale » et nous ne reviendrons pas par la suite sur les mécanismes qui y sont montrés. De plus, étant donné que de nombreuses indications ont déjà été données auparavant, nous ne nous attarderons pas trop sur la réalisation de chaque tâche, mais nous serons plutôt attentifs à l'enchaînement des tâches entre elles, et surtout à leurs conséquences sur l'interprétabilité des données.

Reconnaissance de terrain

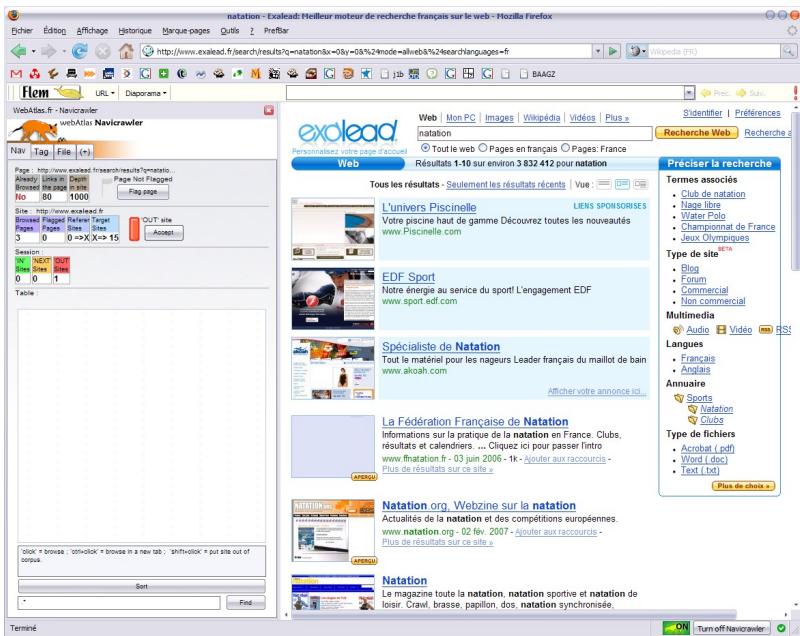
Un usage simple et pourtant souvent négligé du Navicrawler repose sur le simple fait de l'employer « pour voir » comment se présente le web à un certain endroit. Il n'y a pas d'objectif à proprement parler si ce n'est de constituer quelques repères pour la suite.

Nous allons illustrer cette démarche avec le thème de la natation. Les quelques questions auxquelles nous voulons répondre sont les suivantes :

- Quels sont les sites importants
- Est-ce que le domaine « fait communauté » ou pas
- Quels sont ses éventuels sous-domaines

Pour commencer nous n'avons pas de points de départ (bien qu'il soit normalement recommandé de se procurer des URLs auprès d'experts du domaine). Nous allons donc utiliser un moteur de recherche. Cette fois-ci, ce sera Exalead.

- Activons le Navicrawler et rendons-nous sur www.exalead.fr puis entrons la requête « Natation ».
- Il faut écarter le site Exalead lui-même pour qu'il n'apparaisse pas dans le graphe. En effet nous chercherons à voir si les différents résultats sont reliés entre eux, et il ne faut pas qu'Exalead apparaisse comme lien indirect.



- Nous ouvrons dans de nouveaux onglets les URLs qui nous paraissent pertinentes : nous éliminons les URLs publicitaires ou trop décalées par rapport au domaine (par exemple les boutiques qui vendent *entre autres* des articles de natation). Nous cherchons les **sites dédiés** à la natation. Pour ouvrir les liens dans de nouveaux onglets, il faut garder la touche « Ctrl » enfoncée au moment de cliquer sur le lien (si cette action ouvre l'URL dans une nouvelle fenêtre, alors il faut paramétrier Firefox dans le menu Outils>Options...). Nous balayons les deux premières pages de résultats et ouvrons donc une petite vingtaine d'onglets.
- Rien qu'à partir de ce travail nous pouvons déjà remarquer que quelques sites sont apparemment des portails dédiés à la natation, tandis qu'un certain nombre d'entre eux sont des sites de clubs locaux de natation. Remarquons aussi que la « natation synchronisée » est mentionnée plusieurs fois. Nous allons créer des libellés pour ces trois catégories afin de les retrouver plus tard.



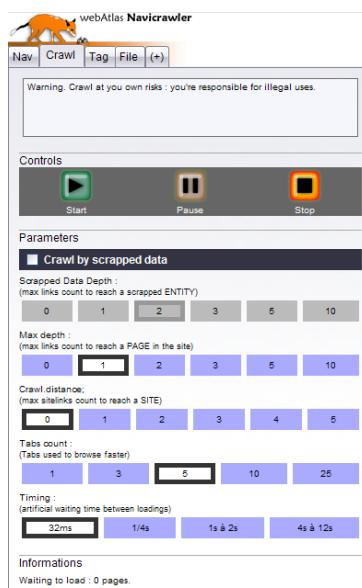
- Nous fermons maintenant l'onglet contenant Exalead et nous regardons chaque autre onglet ouvert pour vérifier que le résultat est bien pertinent (il y a toujours un risque que les liens soient cassés ou que le moteur ne soit pas parfaitement pertinent). Pendant ce balayage, nous classons les sites pour nos trois libellés. Le but est d'aller vite : lorsque le classement n'est pas

évident, nous ne classons pas le site. Attention, nous nous contenterons de passer d'onglet en onglet mais nous ne les fermons pas.

- Nous allons ensuite effectuer un petit crawl à l'intérieur de ces sites, pour cette raison nous fermons deux sites qui seraient embêtants : un forum (nous conseillons de ne pas les crawler sans d'abord comprendre comment ils sont construits) et un site rempli de publicité qui ouvre sans-cesse des pop-ups. NB : les sites fermés sont bien incorporés mais ils ne seront pas crawlés.

Pour faire le crawl, nous allons dans l'onglet dédié du Navicrawler. Si cet onglet n'apparaît pas dans le Navicrawler, vous devez l'activer dans l'onglet « (+) ». Le paramétrage de crawler est le suivant :

- Données heuristiques : non (nous faisons un crawl « normal »).
- Profondeur : 1 (on veut aller assez vite et c'est suffisant)
- Distance : 0 (on reste dans les sites)
- Nombre d'onglets : 5
- Temps d'attente : 35 millisecondes



- On lance le crawl en appuyant sur le bouton vert. Le crawl indique alors 367 pages en attente : comme on est en profondeur 1 il n'y aura pas de ressources à crawler qui seront découvertes en cours de route ; nous patientons le temps qu'il n'y ait plus de pages en attente. Pendant le crawl,

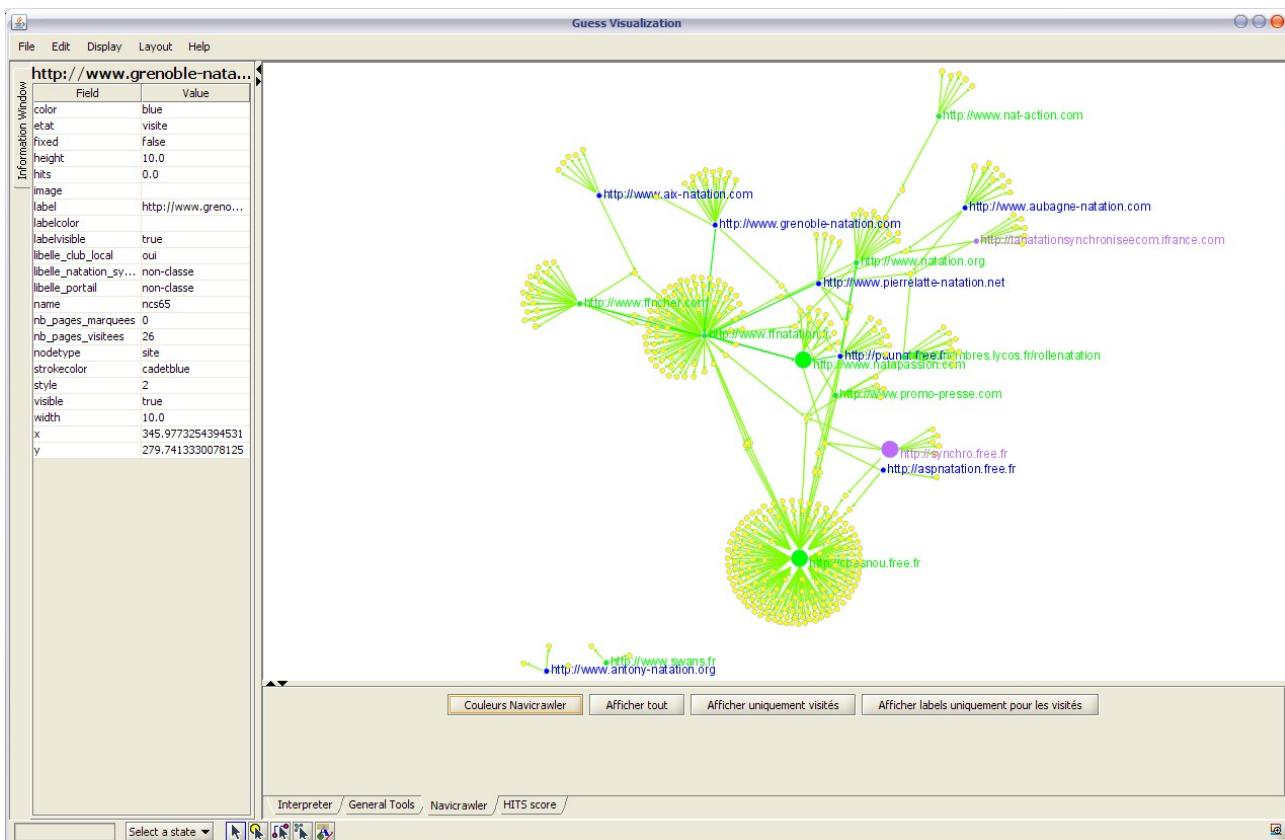
divers pop-ups se sont ouverts qu'il a fallu fermer, ce n'est pas gênant pour les données mais c'est pénible. Huit minutes plus tard, le crawl est effectué. 384 sites voisins ont été détectés.



- Notre exploration rapide est terminée, nous allons regarder ce qu'elle donne. Nous allons donc d'abord sauvegarder le fichier WXSF (pour ne pas perdre les données) puis exporter le fichier GDF pour l'ouvrir avec Guess. Nous ne voulons pas exporter les sites écartés, car nous ne sommes pas dans une démarche de circonscription d'un corpus : il suffit pour cela de décocher la case dans l'onglet « file ». Attention, il faut la laisser cochée pour exporter le WXSF puis la décocher pour l'export du GDF.



- Nous ouvrons maintenant le graphe avec Guess et nous le spatialisons avec l'algorithme GEM puis avec Bin Pack (qui ordonne les composantes non connexes). Puis nous chargeons les scripts et dans l'onglet « Navicrawler » nous colorons le graphe et nous affichons les libellés pour les sites incorporés. De plus, nous mettons les sites portails en gros, les clubs locaux en bleu et les clubs de natation synchronisée en violet (nous avons remarqué pendant l'exploration qu'aucun de nos sites n'était des deux catégories, il n'y a donc pas de croisement à faire).



- Il s'agit maintenant d'observer le graphe pour répondre aux questions que nous nous posons et constituer ainsi quelques repères. Les deux premières observations que l'on peut faire directement sont les suivantes :
 - Deux sites ont beaucoup de liens sortants.
 - Le graphe est globalement connexe mais la plupart des liens se font par l'intermédiaire de sites prochains (en jaune).
- Il faut maintenant regarder de près quels sont tous ces sites. Pour voir le nom d'un site, il suffit passer la souris dessus. Il faut donc balayer avec la souris d'une part les sites qui sont « en étoile » autour des deux sites super-connectés, d'autre part les sites qui font le lien entre les sites incorporés. Ce balayage nous apprend que :
 - Autour du site de la fédération française de natation (celui qui est plus vers le milieu) sont répertoriés des sites de sport génériques, des sites de natation et des sites locaux.
 - Autour du site du bas, cbesou.free.fr, on trouve beaucoup de liens vers des sites de la couche haute (alapage, aufeminin, multmania...), des sites de marques d'articles de sport mais apparemment pas beaucoup de liens vers d'autres sites dédiés à la natation.
 - Les sites qui font le lien entre les différents sites de natation sont pour la plupart des sites de la couche haute : adobe.fr, phpbb.com, google, xiti... On trouve cependant plusieurs sites particulièrement pertinents, dont on

supposera donc qu'il peuvent appartenir au cœur du domaine : la FINA (fédération internationale de natation), centre.ffnatation.fr ; on trouve également des sites dont il faudra arbitrer s'ils appartiennent au cœur ou à la couche haute : www.jeunesse-sports.gouv.fr, www.comite-olympique.asso.fr.

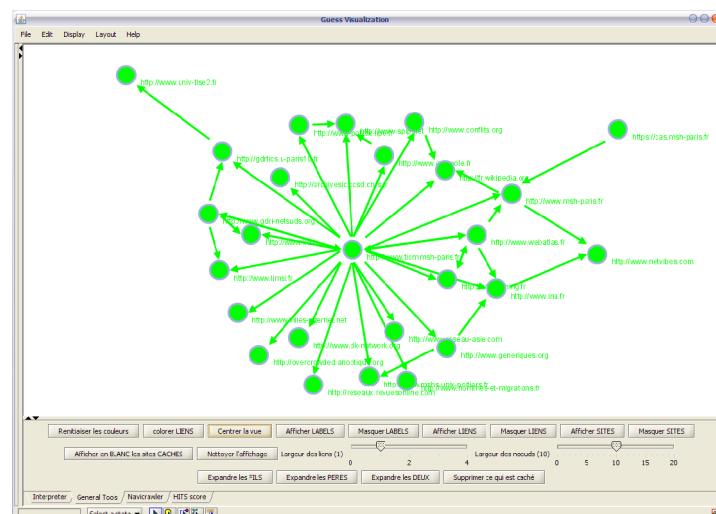
- Sur la base de ces observations nous pouvons déjà remarquer que :
 - Le domaine semble relativement disparate, les URLs sont faiblement liées entre elles
 - Le site de la fédération française de natation semble à la fois être une autorité et un portail pour le domaine
 - Le site de la FINA semble être une autorité importante du domaine
 - Les sites institutionnels (fédérations, ministère des sports, comité olympique) jouent un rôle clé dans le domaine
- Des observations sur la distribution et la répartition en quantité des différentes catégories nous permettent aussi de noter que :
 - Les clubs de natation forment une composante importante du domaine
 - La natation synchronisée apparaît au titre du seul sous-domaine thématique revendiqué, mais il ne semble pas exister comme une communauté
- En conclusion, nous proposons les repères suivants pour orienter une éventuelle exploration future :
 - Le domaine ne semble pas être une communauté. Ceci pourrait cependant être le cas, mais seulement si nos points d'entrée donnés par Exalead ne sont pas à l'origine du « liant » communautaire : il se peut que nous n'ayons pas découvert les sites les plus communautaires (s'ils existent). Il faudrait donc pour en avoir le cœur net chercher explicitement du côté d'une éventuelle communauté de la natation (requête « blog natation », ou « nageurs », ou encore « piscine communauté », « aime nager » etc.)
 - Dans le même ordre d'idée, la présence de l'institutionnel est très forte, soit que la communauté se fonde ailleurs, soit qu'elle n'existe pas. En revanche il est probable qu'un domaine des fédérations de natation à l'international, voire à des échelles plus locales, existe. Ce ne serait cependant pas une communauté à proprement parler mais d'un ensemble d'acteurs « officiels » et de la cristallisation sur le web de leurs liens professionnels
 - On remarque la présence de marques comme Arena ou d'autres. Comme dans d'autres communautés liées à des biens de consommation il pourrait être utile de traquer les forums de discussion autour de ces produits.

- Plusieurs sites nous semblent occuper une position stratégique dans le domaine, ces sites pourraient être de nouveaux points de départ pour une exploration complète du domaine. Parmi ceux-ci on ne retiendra pas les clubs locaux, qui semblent entretenir plus de lien avec d'autres acteurs dans leur localité qu'avec leurs équivalents ailleurs (ce qui est cela dit très classique). Voici donc les sites qui nous semblent majeurs :
 - www.natapassion.com (portail, position centrale)
 - www.ffnatation.fr (officiel, assez central aussi)
 - www.natation.org (lié à d'autres sites majeurs)
 - synchro.free.fr (portail dédié à la natation synchronisée, bien lié)
 - membres.lycos.fr/rollenatation (bien lié aussi)

Étudier le voisinage d'un site

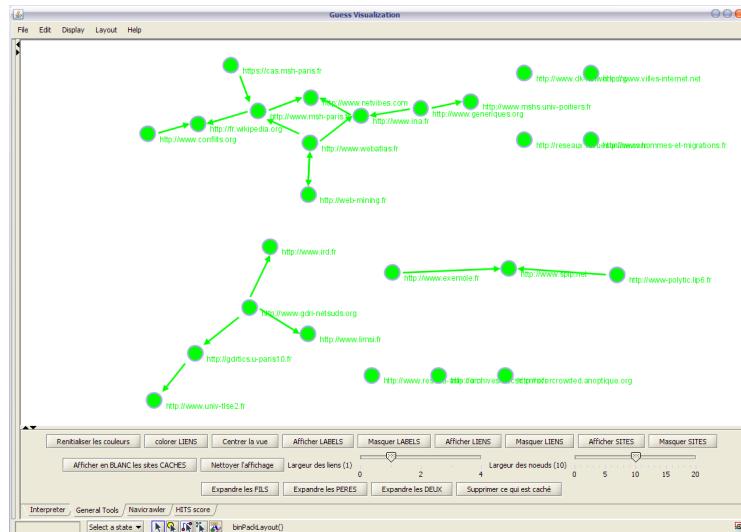
Lorsqu'on s'intéresse à un site en particulier, on peut faire une carte de son contexte, son « environnement ». Le but est d'évaluer ce qui constitue cet environnement : classer les ressources (avec des libellés) et en faire le bilan, observer les regroupements éventuels par liens hypertextes. Comme exemple nous allons explorer le voisinage du site « TIC-Migrations ».

- Le Navicrawler est allumé puis on charge l'url www.ticm.msh-paris.fr et on lance un crawl en profondeur 1 et distance 1. On exporte le graphe (GDF) uniquement des sites visités et on l'ouvre dans Guess, puis on le spatialise avec GEM. On charge les scripts et on applique les couleurs Navicrawler et on affiche les libellés. Voici le résultat :



- Comme on s'y attend, le site TIC Migrations se trouve au centre. On peut néanmoins remarquer quelques sites qu'il ne pointe pas directement. Ces sites sont en réalité pointés (puisque autrement ils n'apparaîtraient pas) mais le Navicrawler n'a pas capitalisé les liens : ceci est sans doute dû à des redirections non gérées. On remarquera aussi que certaines ressources sont

liées entre elles autrement que par le site TIC Migrations, tandis que d'autres ne le sont pas. Pour que ce soit plus clair nous allons supprimer le site central et respatialiser pour observer ces regroupements. On supprime le site avec le menu déroulant qui apparaît lorsqu'on fait un clic-droit sur ce noeud puis on utilise les layouts GEM et Bin Pack.



- Sept sites ne sont pas liés du tout. En dehors de ceux-ci, on observe trois composantes distinctes. Pourquoi ces sites sont-ils liés ? En balayant les noeuds, on peut tirer quelques conclusions :
 - Le paquet du haut est constitué de quelques sites de la couche haute, ainsi que de sites liés à la Fondation Maison des Sciences de l'Homme, et différents partenaires scientifiques et techniques de TIC Migrations (INA, WebAtlas...)
 - Le paquet du bas est constitué de laboratoires ayant pour point commun le développement (non pas au sens informatique, mais au sens international, le développement d'un pays).
 - Le paquet de droite, le plus petit, n'est pas significatif car seul SPIP fait le lien (et il appartient à la couche haute).
- On retrouve donc bien la distinction entre deux volets des activités du programme TIC Migrations : la recherche techno-méthodologique et la recherche sur (ou pour) le développement (relativement au domaine des migrations en tant que tel). Les sites non liés entre eux sont tour à tour orientés technique, migrations ou revues scientifiques.
- On pourrait étendre l'étude en effectuant un crawl avec plus de profondeur (pour récupérer plus de liens dans le site TIC Migrations comme dans les sites autour de lui), ou raffiner l'analyse en utilisant les libellés. Pour aller jusqu'au bout de ce travail et produire une cartographie complète, il serait également utile d'éliminer la couche haute d'emblée. (Wikipedia, Spip etc.)

Une dernière remarque : cette méthode est dite « **focus** » car on s'intéresse à un aspect du web en particulier, un site web donné. Les analyses « focus » n'ont

pas pour but de décrire en tant que tel un domaine, mais elles en sont un bon complément (qu'elles interviennent avant, pendant ou après). La principale différence méthodologique tient au statut du corpus produit. Ici nous n'utilisons pas les caractéristiques incorporé/prochain/écarté pour circonscrire un corpus légitime mais simplement pour collecter une information définie à l'avance (un site et son voisinage). L'avantage de ce type de méthode est de permettre une utilisation plus libre du crawler. En effet, dans les explorations de domaines il est formellement interdit de laisser le crawler incorporer des sites de lui-même, sans quoi on ne peut plus différencier la sélection manuelle de la sélection automatique, et le corpus entier perd sa légitimité. Ici l'interprétation repose sur ce que le crawler a collecté et pour cette raison on peut utiliser des distances supérieures à zéro. Les méthodes focus sont donc :

- déployées pour explorer un objet précis permettant de **définir à l'avance un terrain fermé**,
- en utilisant au besoin les fonctions de crawl en distance supérieure à 0,
- et **l'interprétation se base sur les caractéristiques du terrain prédefini.**

Les méthodes focus permettent différents autres scénarios :

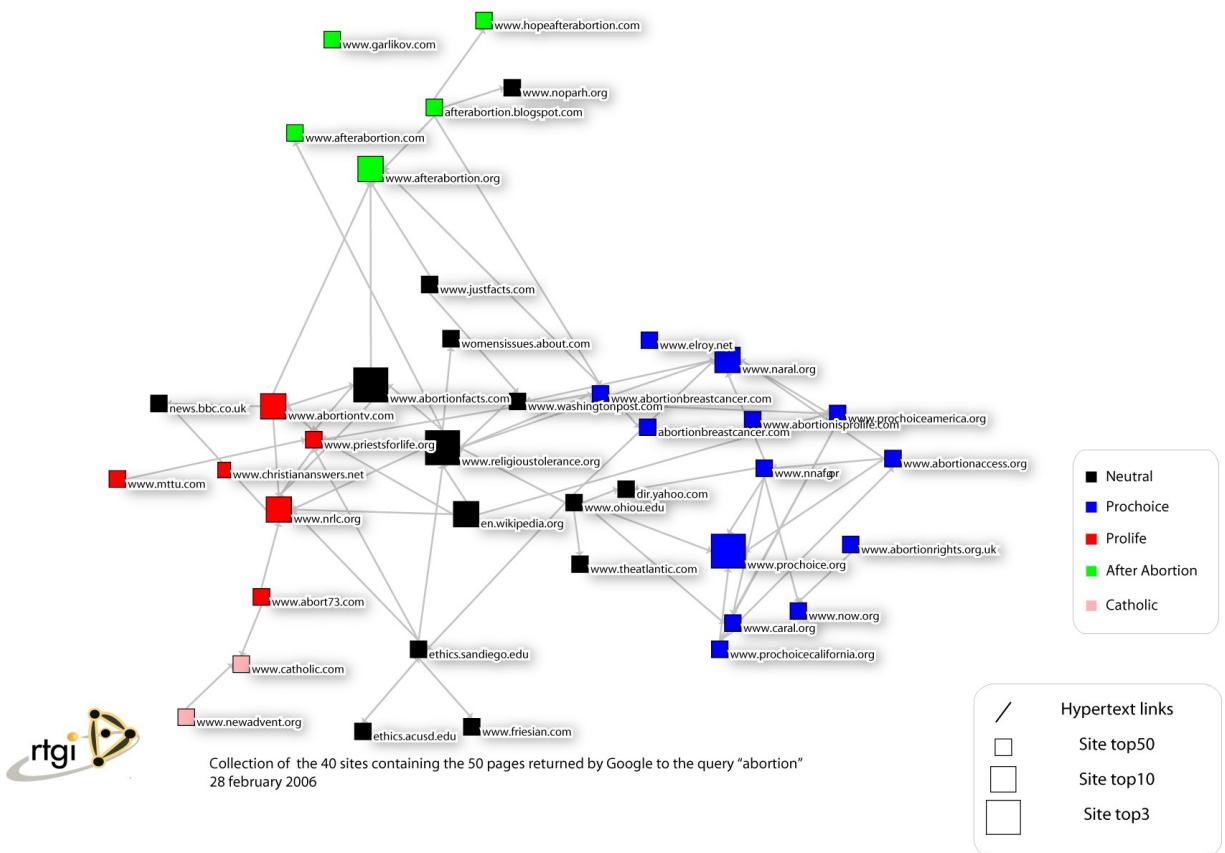
- Crawler deux sites pour savoir s'ils sont connectés
- Crawler un ensemble de sites pour savoir si dans leur voisinage on retrouve d'autres sites attendus par ailleurs. On emploie cette méthode lorsqu'on dispose d'une liste d'URLs dont on veut tester la cohérence. On sépare cette liste en deux aléatoirement pour voir si les sites d'un paquet renvoient bien aux sites du second paquet. On effectue ce travail plusieurs fois de suite, à la manière d'un « sonar ». Cette méthode a l'avantage de ne pas nécessiter de manipulation de graphes (un tableau est suffisant).
- Crawler un ensemble de sites qui pointent tous vers un même site pour savoir s'ils sont reliés entre eux. On cherche ainsi à savoir si le fait de pointer vers une autorité, le CNRS par exemple, est le signe d'une appartenance communautaire ou pas. Le résultat pourra nous aider à classer ce site dans le cœur ou dans la couche haute.

Analyser la réponse d'un moteur de recherche

Cette méthode est particulièrement utile lorsqu'on cherche des informations sur un domaine polémique, et en particulier lorsqu'on anticipe des communautés compétitives. L'exemple prototypique est la controverse autour de l'avortement aux USA, opposant « pro-life » et « pro-choice ». Mais il peut aussi s'agir d'analyser quel profil de site est le mieux placé dans les résultats des moteurs pour une requête stratégique.

- Ouvrir le Navicrawler et accéder à un moteur de recherche, puis taper la requête en question. On écarte le site du moteur de recherche pour qu'il ne trouble pas la carte qui sera élaborée au final.
- On crée un groupe de libellés contenant les différentes catégories attendues (ou on crée ce groupe de libellé en cours de route).
- Pour que la carte soit significative il faut analyser au moins les 50 premiers résultats. Dans ce cas on créera des libellés correspondant à la première apparition d'un site dans la liste (rappelons qu'un site peut apparaître plusieurs fois dans les résultats) :
 - Top 3 (certains internautes ne vont pas au delà)
 - Top 10 (la plupart des internautes vont rarement au delà)
 - Top 50 (si on analyse une grande quantité de résultats)
- Chaque résultat est alors visité et classé. Pour chaque site ouvert on naviguera au moins dans une dizaine de pages en cherchant d'autres liens, dans les pages de partenaires ou dans les pages de liens etc. Nous déconseillons l'usage du crawler, si néanmoins il devait être utilisé (en distance zéro bien évidemment) cette phase serait effectuée à la fin, car le crawl nécessite de fermer les pages qui ne sont pas des points d'entrée et donc de fermer la page du moteur, ce qui est embêtant lorsqu'on est en train de classer les pages.
- Le graphe est alors sauvegardé (uniquement les sites incorporés), puis ouvert avec Guess et spatialisé. On effectue également les traitements suivants :
 - On colore les sites selon les catégories. Attention, les couleurs ont une symbolique et il faut parfois les choisir avec soin. (par exemple le parti socialiste devrait être en rose, l'UMP en bleu etc.)
 - On attribue aux noeuds une taille d'autant plus grande qu'ils arrivent tôt dans les résultats. On s'appuie pour cela sur les libellés « top 3 », « top 10 » etc.
 - L'interprétation repose sur la situation respective des différentes communautés. (sont-elles liées entre elles ? Lesquelles sont les plus liées avec lesquelles ? etc.)

Voici un exemple de ce que donne ce type de travail, dans lequel on observe nettement l'opposition entre pro-choice, pro-life, la situation des sites catholiques et le domaine inattendu « after abortion » :



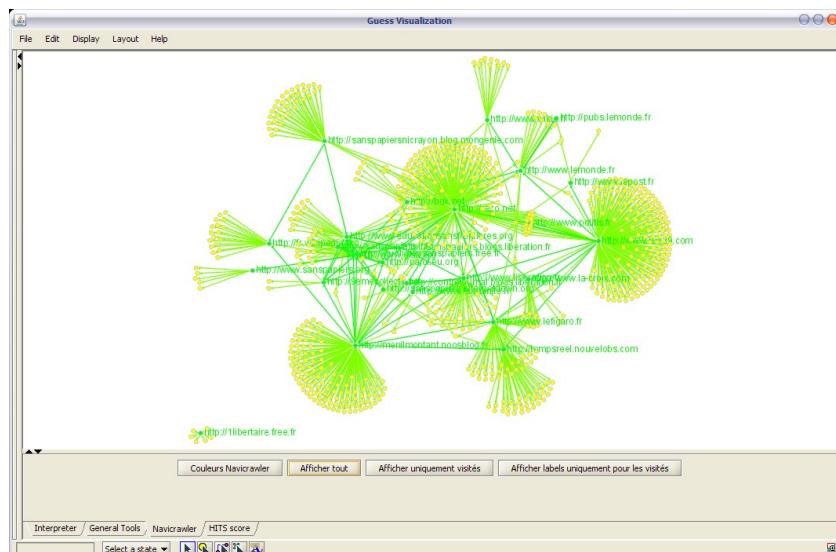
Déetecter les centres

Une tâche relativement simple mais utile consiste en la recherche explicite du centre d'une communauté. Le principal trait de cette méthode est l'aller-retour nécessaire entre l'exploration et la visualisation de graphes. Notez bien ceci : la méthode proposée ici vous permettra de rechercher le coeur le plus proche, mais ceci ne signifie pas que ce coeur sera le coeur du domaine correspondant au domaine attendu ; en effet si le domaine attendu n'est pas agrégé sur le web, on trouvera le coeur de l'agrégat le plus proche. Identifier ce phénomène est justement l'un des résultats attendus de cette méthode : lorsqu'on souhaite évaluer rapidement un domaine sans l'explorer entièrement, il est stratégique de déterminer que coeur « teinte » les ressources. Le cas typique est la présence d'une certaine thématique à un niveau non agrégé (couches plus basses) qui est « phagocyté » par un coeur d'une thématique différente. Citons comme exemple déjà étudié le cas des sites communistes dans le web francophone, qui sont dépendants, dans la géographie de l'information et donc pour l'accès par les internautes, d'un coeur de sites altermondialistes.

Le principe de cette méthode est reposé sur la découverte de nouvelles ressources, comme on la vue dans la partie « reconnaissance de terrain », à ceci près qu'**on se concentre sur les ressources les plus connectées**. On sélectionne donc à chaque étape les ressources les plus connectées pour ne pas

s'éparpiller, jusqu'à une analyse thématique rapide de ce que contient la zone la plus dense relative à notre domaine.

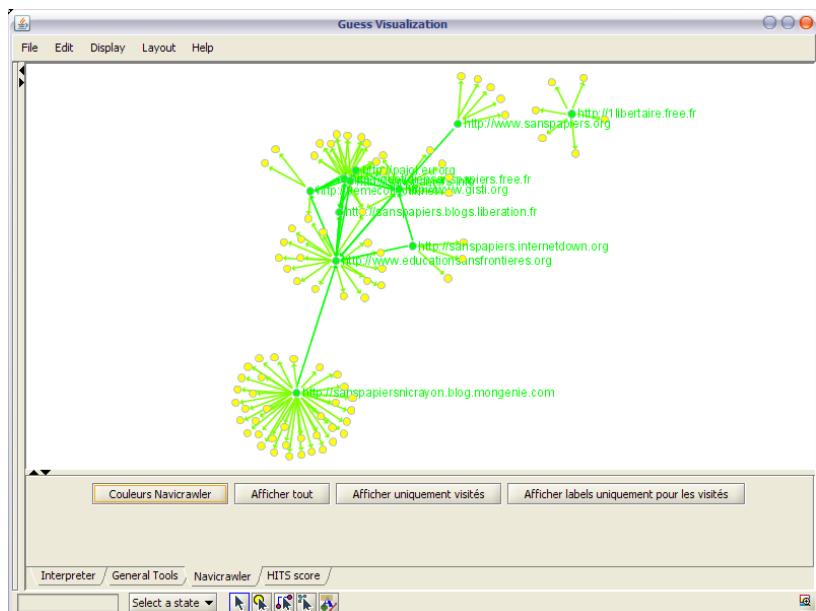
- Nous allons chercher le cœur le plus proche de la thématique « sans papiers ». Commençons par une requête Google et visitons rapidement les sites des trois premières pages de résultats. Comme d'habitude on écarte le site Google. On exporte le graphe des sites en gdf, mais sans les sites écartés (on décoche la case au moment de l'export). Puis on ouvre le graphe dans Guess, on le spatialise, et on lui applique les couleurs Navicrawler.



- On alterne l'affichage uniquement des sites incorporés et l'affichage de tous les sites, pour savoir si le domaine est connexe. Dans notre cas le domaine est clairement connexe. On balaye les sites incorporés pour vérifier qu'ils appartiennent bien au domaine. Plusieurs cas peuvent se produire :
 - Soit les sites ne sont pas vraiment connexes, ce qui signifie que le domaine ne dispose pas d'un cœur. Lorsque c'est le cas, il faut sélectionner les sites les plus connectés et explorer leur voisinage en priorité par les sites prochains cités plusieurs fois.
 - Soit les sites sont assez densément connectés mais une partie des connections ne correspondent pas au domaine, auquel cas il faut éliminer ces sites pour trouver un cœur propre au domaine
 - Soit les sites sont assez densément connectés et ils appartiennent tous au domaine : dans ce cas le domaine dispose d'un cœur et un peu d'exploration permettra d'affiner la détection du cœur mais globalement le résultat sera d'ors et déjà positif.
- Ici nous nous trouvons dans le second cas. En effet de nombreux sites sont des sites d'information, qui n'appartiennent pas spécifiquement au domaine. De tels sites appartiennent à la couche haute, et ils nous empêchent de voir le cœur. C'est le cas dans de nombreux domaines : les sites d'information (journaux, Wikipedia), les répertoires de contenus juridiques ou

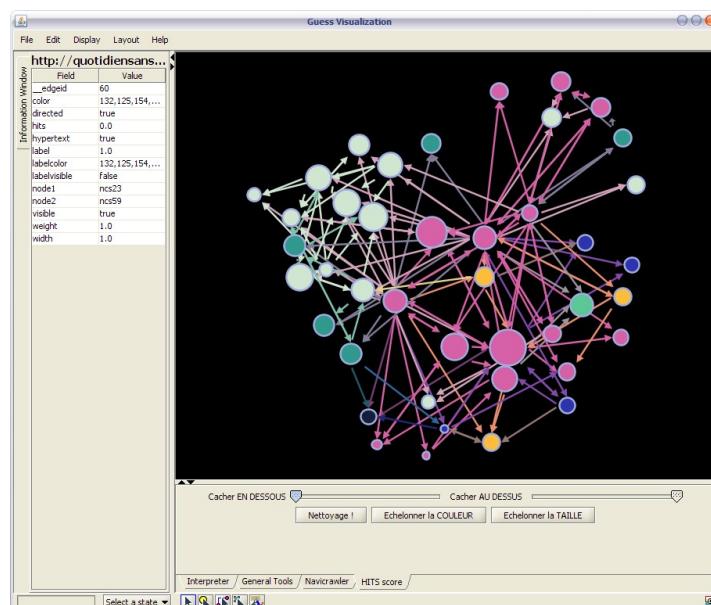
scientifiques, sont souvent massivement cités dans des thématiques qui s'appuient sur leur contenu à titre de référence. Il vous appartient de trancher leur appartenance à la couche haute ou au cœur de votre domaine. Nous préconisons de les classer dans la couche haute car ils sont si génériques qu'ils nous empêchent de voir si oui ou non une agrégation se produit au niveau du domaine en lui-même.

- Les sites qui n'appartiennent pas au domaine (soit qu'ils soient de la couche haute soit qu'ils soient d'autres domaines) sont écartés directement dans le Navicrawler. La technique est simple : on affiche la liste des sites incorporés en cliquant sur la case verte dans l'onglet « Nav », puis on visite les sites les uns après les autres en écartant ceux qu'il faut. On peut aller plus vite en écartant les sites sans les visiter : il suffit d'appuyer sur la touche majuscule du clavier et de cliquer dans la liste sur le site qu'on veut écarter. C'est la technique que nous utilisons car les adresses à éliminer sont des adresses bien reconnaissables. (Libération, Le Figaro, l'Humanité, etc.) Puis on réouvre de la même façon le graphe exporté, sans les sites écartés, dans Guess.



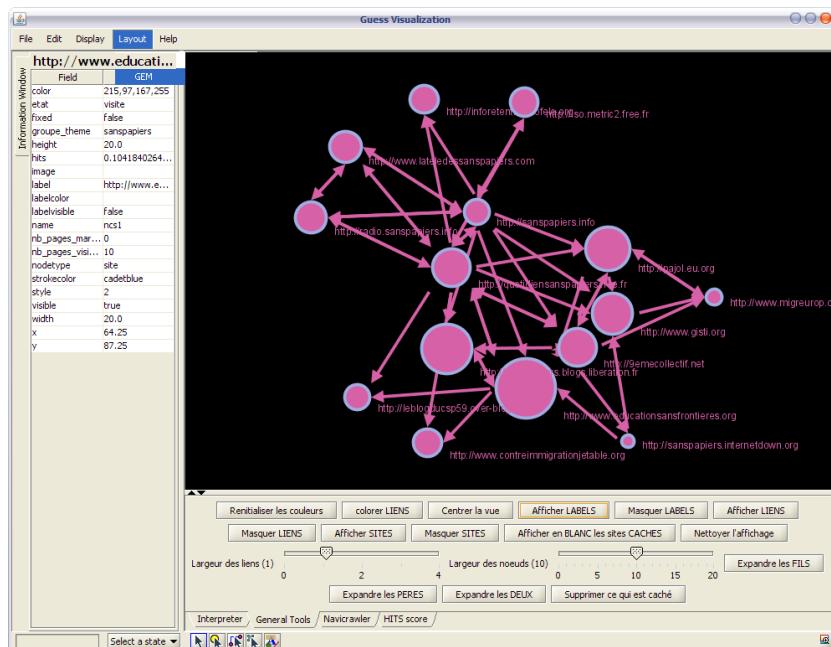
- Un ensemble de sites apparaît nettement interconnecté : c'est probablement le cœur ou une partie du cœur. Que les sites soient fortement connectés ou pas, la suite de la technique est la suivante :
 - on recommence une session Navicrawler à partir des sites du paquet central, indice du cœur. Cette fois on visite ces sites attentivement pour bien collecter tous les liens.
 - On crée un libellé propre au domaine (ici « sans-papiers ») dans lequel on classe les sites actuellement incorporés.
 - Puis on élimine les sites prochains cités une fois seulement (dans notre cas il ne reste que 37 sites prochains pour 9 sites incorporés).

- Enfin on classe les sites prochains restants dans de nouveaux libellés, en écartant ceux qui sont de la couche haute et dont la présence n'est pas pertinente. L'ensemble de ces libellés doit être dans un même groupe. Dans notre cas nous avons écarté les sites anglophones, et diverses scories comme des intranets etc. Les libellés que nous avons créés pour décrire le contenu sont les suivants :
 - Divers solidarité
 - Lutte contre le racisme
 - Solidarité Internationale
 - Juridique
 - Presse/Médias
 - Sans Papiers
- Attention, il est utile de régulièrement éliminer les sites prochains cités une fois seulement au cours de cette exploration, car comme on visite de nouveaux sites, de nouveaux sites voisins apparaissent. Cependant, de nouveaux sites prochains cités deux fois ou plus apparaissent. A un moment donné on se rend compte que ces sites appartiennent à peu près tous à la couche haute. C'est le moment où on stoppe l'exploration. Dans notre cas nous avons incorporé 42 sites, écarté 15 d'entre eux et il reste encore 50 sites prochains qu'on ne visitera pas.
- Encore une fois on exporte le graphe, cette fois-ci uniquement des sites incorporés. On les ouvre dans Guess et on visualise le graphe. Le graphe est coloré d'après les libellés du groupe que nous avons créé plus haut, et la taille est affectée d'après la connectivité (onglet « HITS » dans les scripts chargés dans Guess).

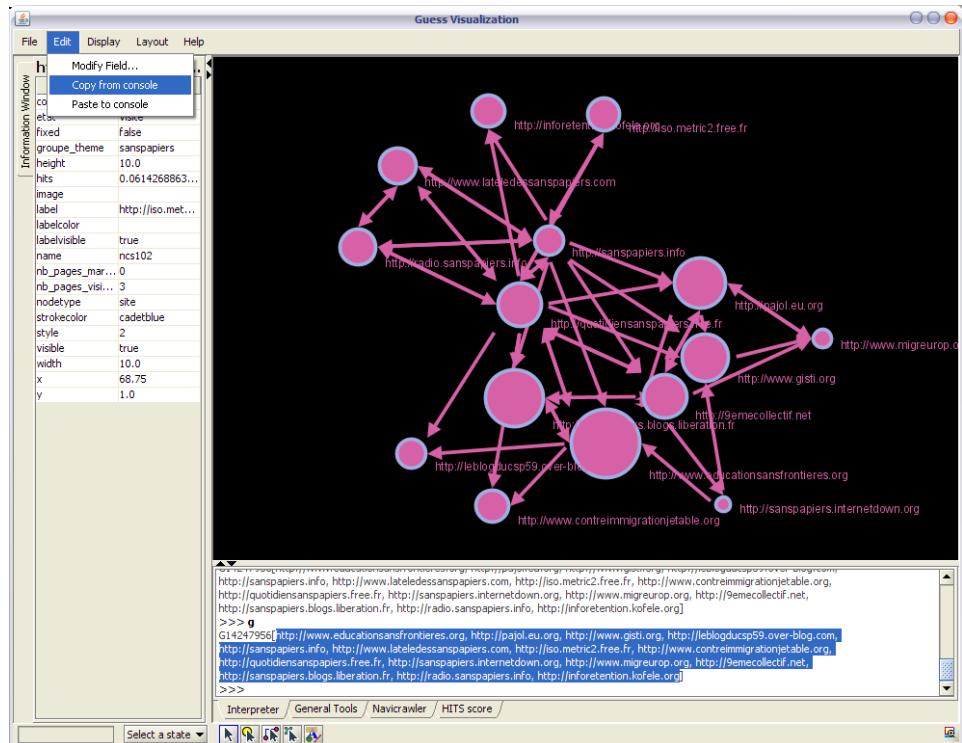


- Sur la gauche du graphe, en gris clair, on peut voir une nette zone

d'agrégation. Ce sont les sites Presse/Médias. En rose, au centre, on retrouve comme on s'y attend les sites Sans-papiers. Les autres catégories ne se regroupent pas vraiment. C'est bon signe ! Notre conclusion est que le domaine des sans-papiers dispose bien d'un cœur et qu'il est, dans la couche haute, fortement lié aux sites de presse. Remarquons par ailleurs que les sites Sans-papiers disposent de gros scores de HITS pour la plupart : ils sont bien centraux dans leur environnement. En supprimant les sites qui ne sont pas Sans-papiers et en respatialisant on fait apparaître uniquement ce que nous considérerons comme le cœur du domaine.



- En tapant simplement la lettre « g » puis « entrée » dans le champ de commandes de Guess, la liste des noeuds sera affichée. On peut alors la sélectionner puis la copier depuis le menu « Edit ». Vous récupérez ainsi la liste des sites du cœur, que voici.



- <http://www.educationsansfrontieres.org>
- <http://pajol.eu.org>
- <http://www.gisti.org>
- <http://leblogducsp59.over-blog.com>
- <http://sanspapiers.info>
- <http://www.lateledessanspapiers.com>
- <http://iso.metric2.free.fr>
- <http://www.contreimmigrationjetable.org>
- <http://quotidiensanspapiers.free.fr>
- <http://sanspapiers.internetdown.org>
- <http://www.migreeurop.org>
- <http://9emecollectif.net>
- <http://sanspapiers.blogs.liberation.fr>
- <http://radio.sanspapiers.info>
- <http://inforetention.kofele.org>

Circonscrire et analyser un domaine

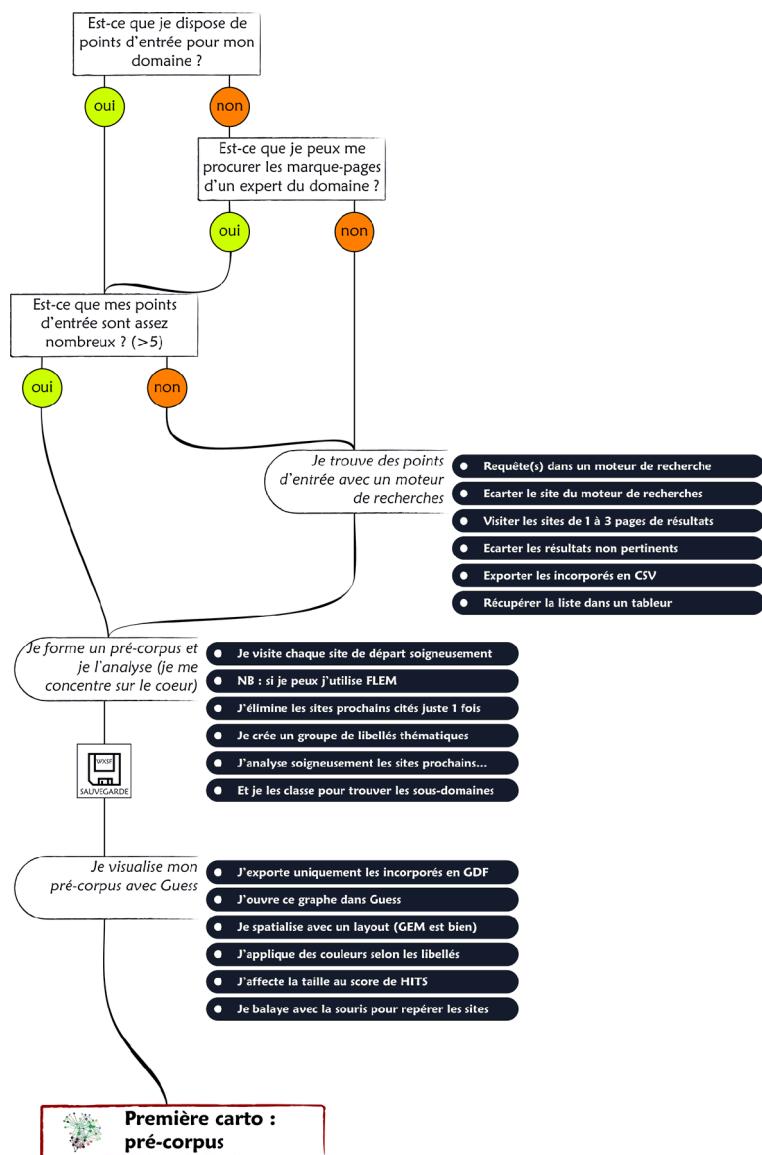
L'analyse complète d'un domaine requiert plusieurs étapes que nous allons lister les unes après les autres. L'essentiel des manipulations nécessaires ont déjà été détaillées plus haut, nous allons donc nous concentrer sur l'enchaînement des étapes et les différentes questions qui se posent au long de l'analyse.

Les différentes étapes que nous allons décrire pas à pas sont les suivantes :

- Collecter et visualiser le pré-corpus
- Interpréter la première carto
- Etendre le corpus
- Affinage et analyse de l'intérieur du corpus
- Analyse de la frontière
- Production de la carto finale

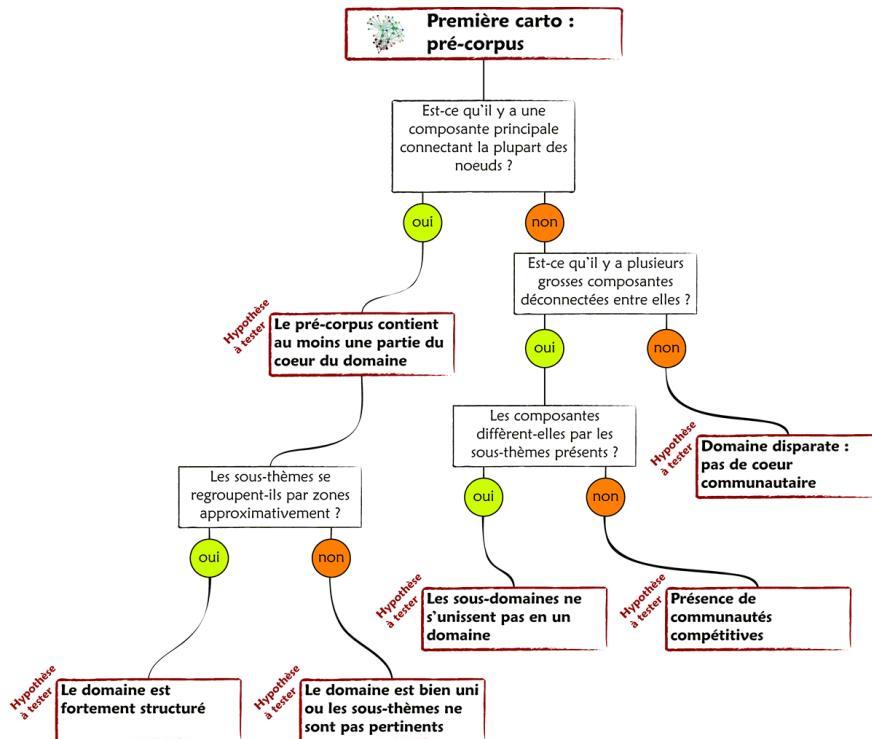
Collecter et visualiser le pré-corpus

La première étape consiste en une exploration du domaine en vue de produire une première cartographie que l'on va analyser pour cadrer la suite des tâches. Il est question ici de bien engager l'exploration pour ne pas perdre trop de temps au démarrage.



Interpréter la première carto

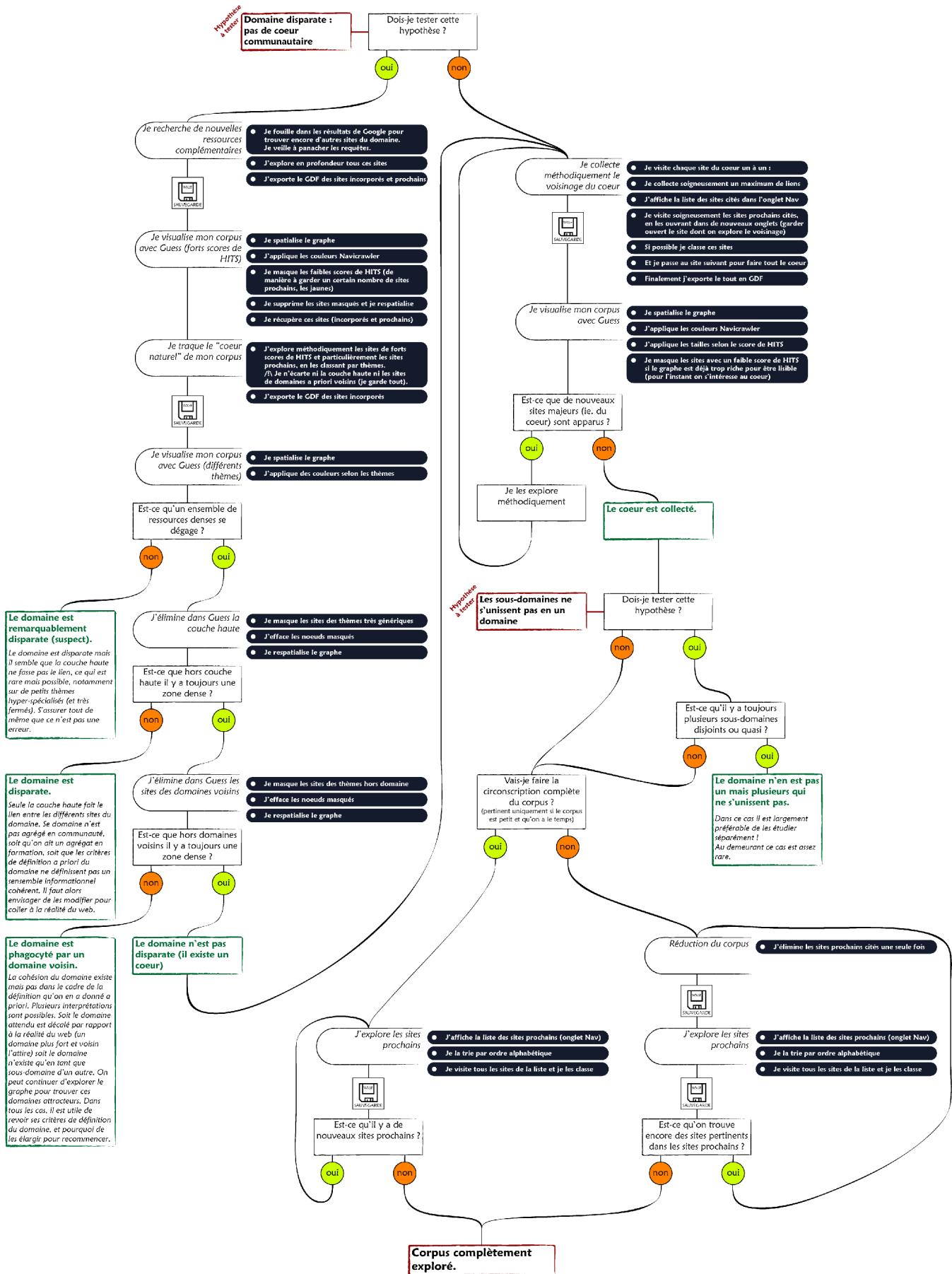
Cette phase ne comporte pas d'opérations techniques à proprement parler. Il est simplement question de bien regarder le graphe et de le comprendre. Mais bien sûr il ne vous est pas interdit de manipuler le graphe pour mieux comprendre sa structure ! Nous cherchons à évaluer la consistance du domaine en observant ses éventuelles zones de concentration et la distribution des thèmes (d'où la nécessité de les renseigner lors de la création du pré-corpus).



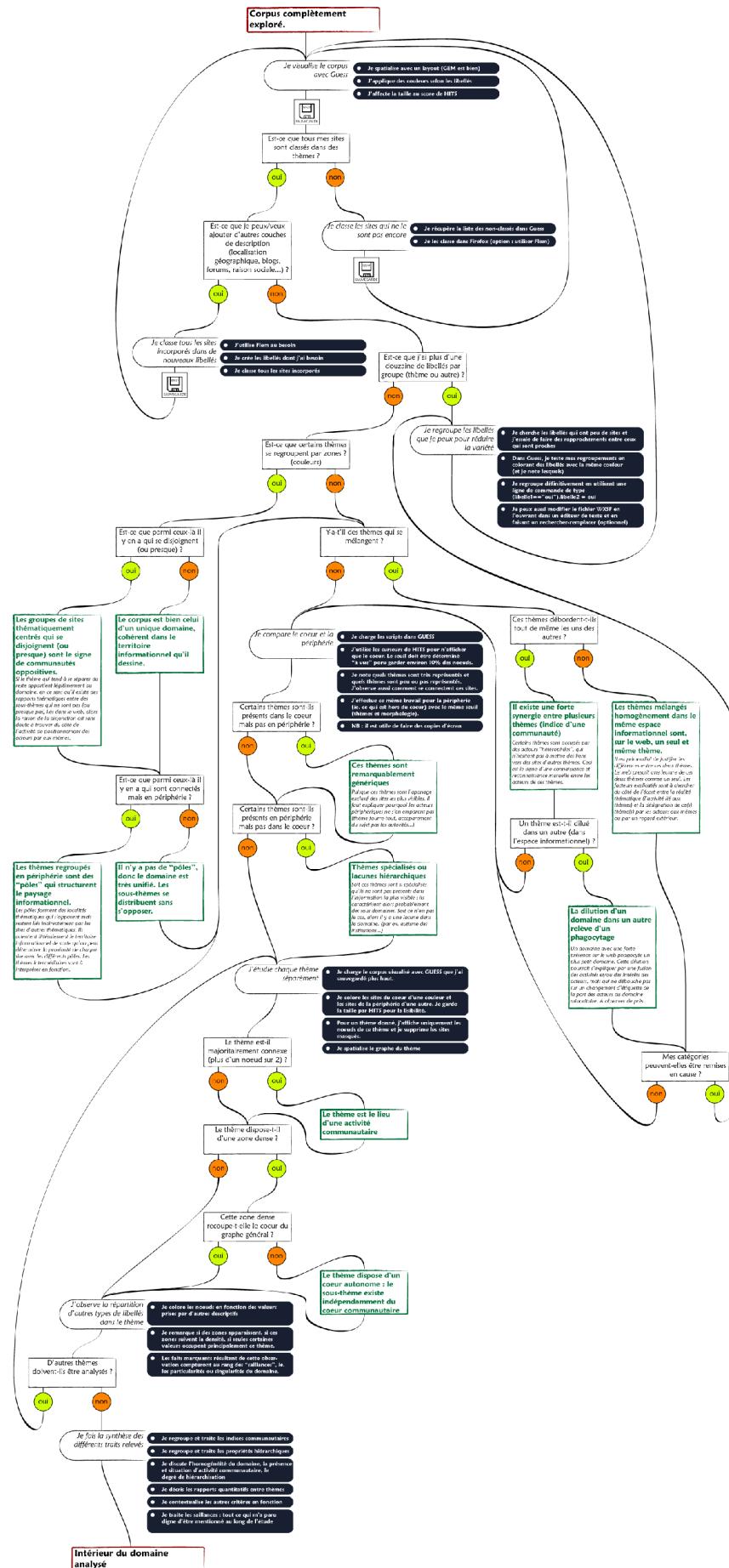
Certaines des hypothèses formulées seront testées explicitement tandis que d'autres seront testées « naturellement » au cours des étapes suivantes. Cependant le fait de formuler ces hypothèses à cette étape, même si elles sont de toute façon éprouvées par la suite, vous aidera à porter un regard plus pertinent sur les données que vous analyserez.

Par ailleurs cette étape peut faire office d'évaluation : c'est le bon moment pour décider de continuer ou d'arrêter l'exploration. Si vous ne voulez pas explorer un domaine non-constitué et qu'il semble ici qu'il n'y a pas ou peu de teneur communautaire, choisissez un autre thème ou élargissez d'emblée vos critères.

Etendre le corpus

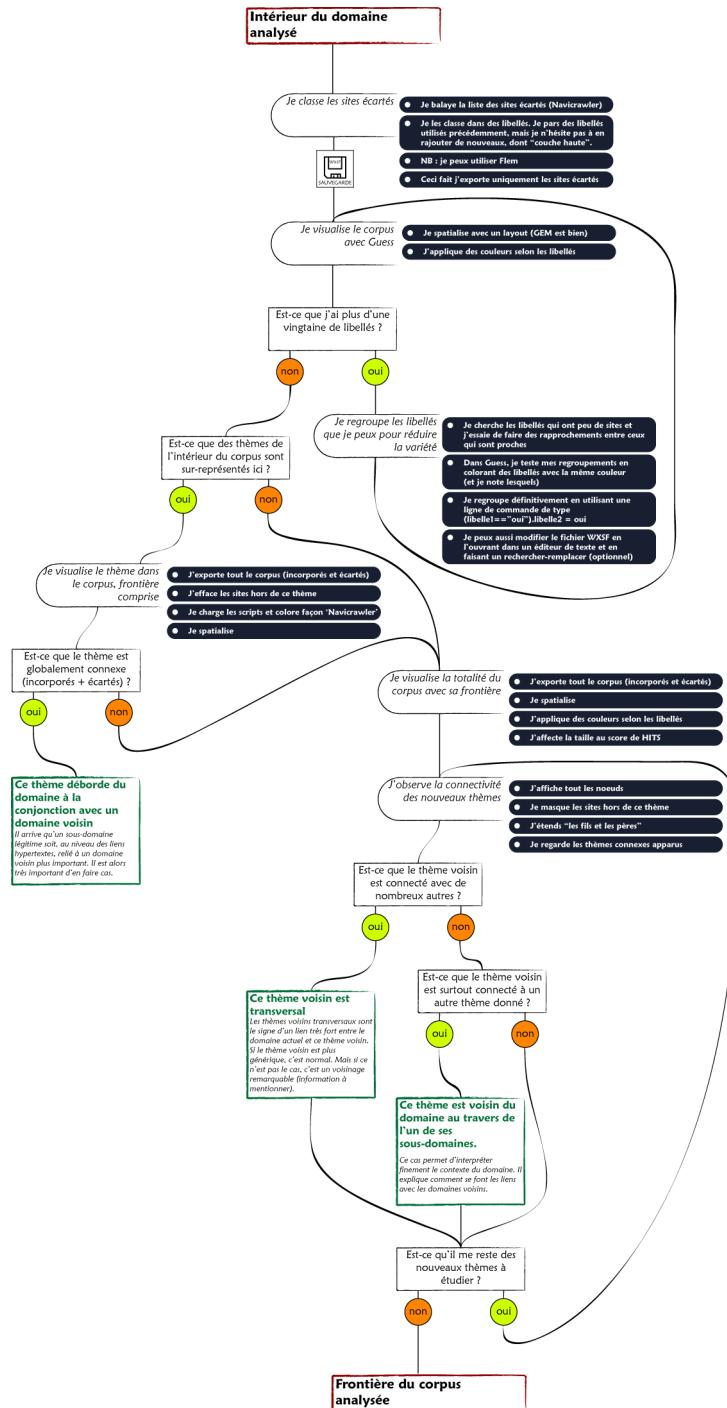


Affinage et analyse de l'intérieur du corpus

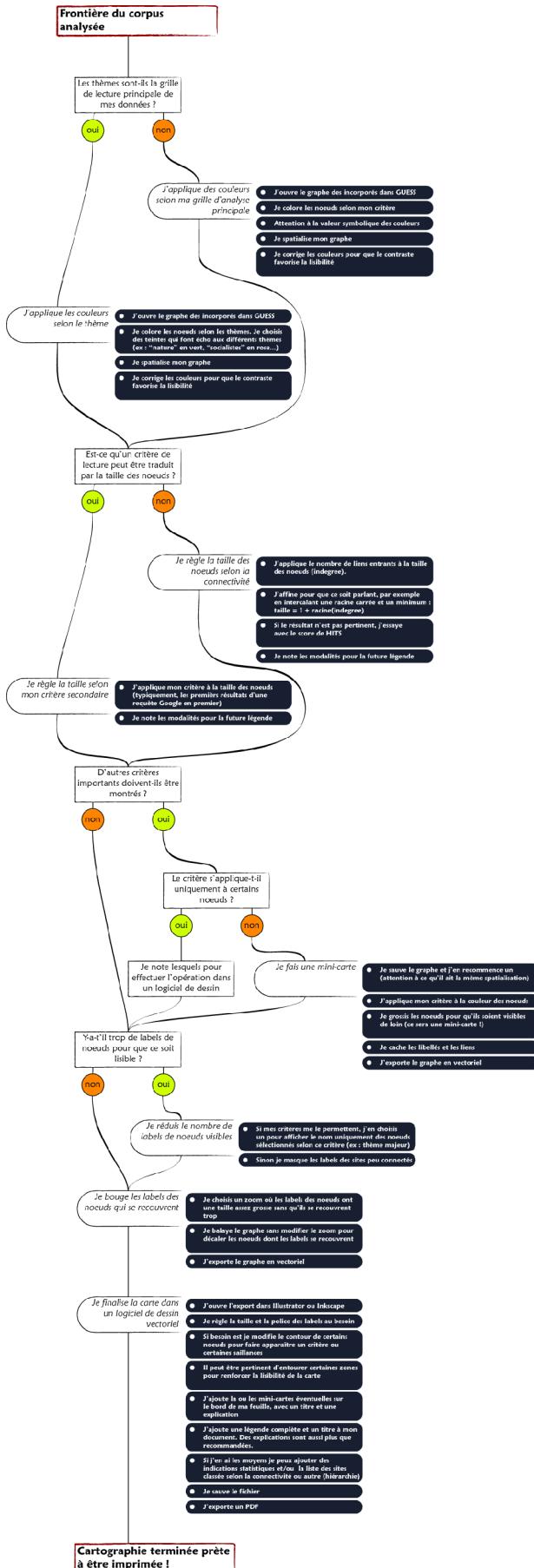


Analyse de la frontière

L'analyse de la frontière est essentielle puisqu'elle permet de contextualiser le domaine. Rappelons que le web est un milieu ouvert, dans lequel il est épineux de « découper » un terrain sans savoir précisément ce que cela implique. L'étude de la frontière permet justement de faire cas de l'environnement du corpus que l'on a déterminé.



Production de la carte finale



Mémo : récapitulatif des principaux conseils

- « Qui se ressemble se connecte » : l'exploration d'un domaine est possible parce que **les ressources autour d'un même thème ont tendance à se lier**. Ce principe guide l'ensemble des méthodes Navicrawler.
- Pour se repérer lors de l'exploration, il faut distinguer la couche haute (les grands sites génériques) du cœur du domaine (ses sites majeurs) et avoir conscience de l'existence d'une nébuleuse autour du cœur, et au-delà de filaments, qui sont plus difficiles à collecter.
- L'exploration doit se faire **en commençant par la couche haute, puis le cœur du domaine, puis la nébuleuse et enfin les filaments**. Une étude rapide se concentre sur le cœur et une partie de la nébuleuse.
- Le principe sites incorporés / prochains / écartés doit être bien compris pour utiliser correctement le Navicrawler. En particulier, les sites écartés n'ont pas de sites prochains. C'est pourquoi ils forment littéralement la frontière du corpus.
- Le Navicrawler permet d'éliminer les sites prochains mineurs pour accélérer l'exploration. Mais il ne faut pas le faire avant d'avoir collecté le cœur pour être sûr de n'avoir oublié aucun site majeur.
- **La description des ressources ne doit pas être confondue avec la sélection des ressources**. Les libellés peuvent être mobilisés pour sélectionner le corpus. Une fois le corpus stabilisé, il faut recommencer la description des ressources de façon indépendante.
- Les libellés s'organisent naturellement en couches de description (raison sociale, thème, géolocalisation...) et en caractéristiques indépendantes (dispose d'un blog, forum, site portail...). **Les couches de description ne doivent pas interférer**.
- La fonction « crawl » permet au Navicrawler de naviguer « tout seul » sur le web. **Utilisez le crawl avec parcimonie !** Vous ne devez jamais laisser le crawl incorporer des sites de lui-même si vous en avez déjà incorporé manuellement (sinon vous ne vous y retrouverez plus).
- Le crawl devra souvent être fait en distance 0. Par ailleurs, il est rare qu'une profondeur supérieure à 2 soit pertinente.
- Utilisez l'export CSV (pour tableur) afin de trier ou de sélectionner vos ressources. L'extension Flem est alors utile pour gérer les listes de sites.
- Pour voir le graphe de sites, exportez les données au format GDF et ouvrez-le avec Guess. Vous pourrez alors spatialiser le graphe, colorer les noeuds selon les libellés, et déterminer une hiérarchie des sites selon leur

connectivité.

- Dans un graphe, observez en priorité les rapports quantitatifs entre différents groupes de sites, et ensuite la façon dont ils sont connectés entre eux.
- La connectivité des ressources majeures entre elles et avec le reste du domaine permet de déterminer un degré d'activité communautaire dans le domaine. Mais ce degré **doit être validé par une observation du contenu des sites eux-mêmes**.
- N'oubliez pas que vous pouvez utiliser le Navicrawler pour des plus petites tâches que l'exploration d'un domaine, par exemple l'analyse des résultats d'une requête Google. Dans ce cas, **déterminez précisément votre terrain** pour ne pas être coincé au moment de l'interprétation.