

# Explorer et Appréhender le web

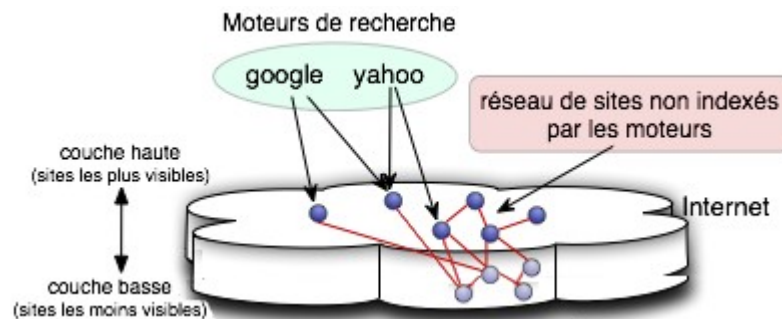
Fabien Pfaënder, Mathieu Jacomy  
Université de Technologie de Compiègne  
Boite Postale 60319  
60603 Compiègne cedex, FRANCE  
+33-3-44-23-52-48  
{fpfaende,jacomyma}@utc.fr

## Introduction

Les systèmes complexes sont des systèmes dont nous n'arrivons pas à nous saisir. La complexité qui les caractérise naît de l'impossibilité à découvrir l'organisation ou la structure de ces systèmes et d'en faire un objet de pensée. Ils se composent en effet d'un grand nombre de données hétérogènes, dynamiques ou non, liées entre elles sans que l'on ne sache comment. Les connaissances que l'on peut faire naître de ces systèmes proviennent en grande partie de la découverte de leur structuration, c'est à dire la façon ou les principes qui lient les données entre elles. Comprendre les systèmes complexes et en extraire l'essence est d'autant plus important que ces systèmes sont de plus en plus nombreux aujourd'hui. Deux raisons principales à cela : d'une part, l'utilisation du numérique permet d'encoder toutes les données quelles qu'elles soient et de les manipuler formellement, indépendamment de leur signification. D'autre part, la démarche scientifique pousse à considérer le maximum de données afin de "coller" au mieux à la réalité du phénomène à observer. Toutes les données contextuelles possibles doivent donc être prises en compte pour expliciter et modéliser le phénomène. Tout le problème consiste à savoir quelles données sont pertinentes et interviennent dans le phénomène et lesquelles ne le sont pas. Aussi, et pour être sûr de ne pas en omettre, il est plus prudent d'en intégrer un maximum au système qui du même coup voit sa complexité augmenter. Le problème de la circonscription du phénomène n'est pas nouveau et les mésopotamiens se posaient déjà la question [Bottero 1987] de la clôture d'un phénomène en considérant un maximum de possibilités dans leur divination déductive, y compris les plus improbables. Ils prévoient par exemple le cas où un mouton aurait 2 foies, mais aussi 3, 4 et ainsi de suite jusqu'à 7 et les significations associées. De la même façon, la théorie du chaos pose pour hypothèse qu'un mouvement d'aile de papillon d'un côté de la planète aura des conséquences désastreuses de l'autre côté [Lorenz 1972] et trouve aujourd'hui une signification toute particulière. On voit bien ici que la tendance consiste à prendre en compte un maximum de données apparentées dans un modèle très large et ouvert. Cela prend de nos jours une ampleur sans précédent avec le besoin impérieux de mieux comprendre le monde qui nous entoure, indispensable pour prévoir notre futur proche.

On trouve donc des systèmes complexes dans tous les domaines. Il en est cependant un, transdisciplinaire et immédiatement accessible, dont les enjeux industriels et scientifiques sont particulièrement importants : le web. Ce dernier est en effet un système complexe très dynamique dont l'explicitation permettrait de tirer efficacement des connaissances pertinentes que ce soit du point de vue documentaire,

social ou de l'organisation de ces documents. le web est une construction qui met en présence des documents hypertextuels liés entre eux. Les documents forment des pages qui, suivant une même unité sémiotique, forment une entité que l'on appelle site Internet (il existe d'autres définitions du site web et cette notion est source de nombreux débats). La distribution de ces sites et documents liés entre eux forme un réseau complexe dont la topologie (distribution des liens) et la distribution du sens suivent des règles qui nous sont, dans une large part, inconnues. La diversité des formats sémiotiques des pages et la liberté de création des liens contribuent à rendre le système hétérogène, augmentant d'autant la difficulté de mettre à jour son organisation globale. Des structures doivent pourtant exister au sein de ce système complexe et les mettre à jour et les manifester est indispensable pour arriver à appréhender le web dans son ensemble et en faire naître de la connaissance. Bien sûr il existe déjà des outils comme les moteurs de recherche (Google™ et Yahoo™ en sont les représentants les plus célèbres) qui permettent de plonger dans la structure de



*Illustration 1: Organisation du web*

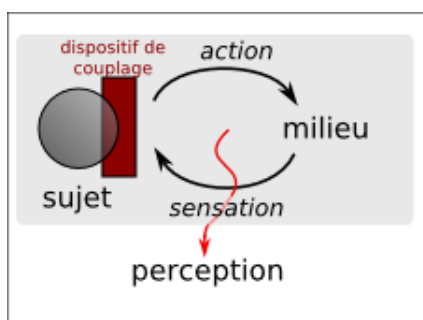
manière ponctuelle mais ils ne permettent pas pour autant d'en rendre compte. Ils sont comme des satellites donnant un accès immédiat aux documents les plus visibles issus d'une requête avec une pertinence relative. Les navigateurs Internet (Firefox, Internet Explorer, etc.) permettent eux aussi un accès aux documents et une très faible spatialisation en utilisant les boutons "avance" et "recul" et plusieurs fenêtres ou onglets. Dans la mesure où ces boutons représentent les liens qui ont été suivis en avant et en arrière de la page actuelle, ils manifestent une spatialisation très locale de la structure du web à cet endroit précis ou d'un chemin particulier. Cela reste très limité et ne permet pas de se construire une représentation [Ghitalla 2003] efficace de la structure de telle ou telle localité d'Internet. Tout au plus cela autorise la création d'un micro-espace mais il est trop immersif pour être efficace. Car c'est là le principal enjeu : proposer un outil qui explore la structure hypertextuelle, en dégage les caractéristiques pour construire un corpus finalisé, avant d'en rendre compte. Le résultat ne sera alors plus une liste de résultats décorrélés incomplète ou un empilement de fenêtres mais une présentation synoptique, un guide, qui rend compte de la structure dévoilée d'Internet et dont le résultat est par conséquent plus pertinent tant du point de vue objectif (les documents sont pertinents) que subjectif (le navigant a prise sur ses résultats et leur organisation). Des outils comme le moteur de recherche KartOO prennent en compte cette dimension et proposent une vue cartographique du résultat des requêtes et une manipulation interactive. Toutefois, il manque toujours un aspect essentiel à ces visualisations qui est l'explicitation de la structure sous-jacente d'Internet. La façon dont le corpus a été analysé n'est pas manifeste dans l'outil. La carte de KartOO par exemple utilise une convention de

hauteur géographique (les lignes de niveaux) sans que cette dernière (qui n'a pas d'existence en soi sur le web) ne soit clairement associée à un critère sur la carte. On ne peut que faire des hypothèses sur sa signification et le résultat apparaît alors comme décorrélé de la réalité en dépit de son intérêt. De plus, les moteurs de recherche ne proposent souvent qu'une sélection de sites qu'ils jugent pertinents suivant certains critères sans jamais explorer la structure d'un domaine précis et en l'analysant complètement.

Afin de construire cet outil, nous proposons une méthode d'exploration et de préhension de l'organisation et de la structure du web qui peut, sous certaines conditions, être mobilisée pour explorer n'importe quel système complexe. Dans un premier temps, nous présenterons la question de la prise, de la trace et de l'appréhension, base théorique qui sous-tend l'ensemble de la démarche. Dans un second temps, nous présenterons les trois étapes de la méthode que sont la récolte de données, leur exploration et enfin leur synthèse pour transmettre les connaissances acquises sur le système.

## Appréhender

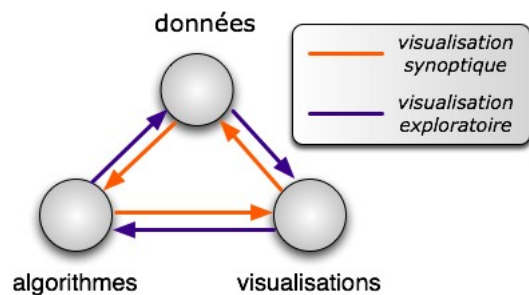
Que signifie comprendre un système complexe comme Internet ? L'un des objectifs est d'en découvrir la structure mais cela ne suffit pas à en faire un objet de pensée. La connaissance de l'organisation permet en effet d'outiller une exploration vers des parties du système qui deviennent accessibles, situées dans un espace, mais cela ne signifie pas pour autant que l'on est en mesure d'en apprécier l'ensemble. La connaissance du système n'est donc dans ce sens que partielle. Comprendre un système complexe dans son sens large signifie pouvoir s'en saisir comme d'un objet, pouvoir l'appréhender. Le système auquel nous avons affaire devient un espace de perception dont on connaît les lois [Lenay 2005]. On sait s'y déplacer et l'on sait comment les éléments qui le composent sont agencés. Il est alors aisé de l'utiliser et d'en tirer les informations que l'on souhaite. Cependant atteindre ce but est éminemment difficile, et ce dans la mesure où le système complexe est ce qu'il est justement parce qu'il n'offre pas de prise directe. Il faut donc trouver la façon d'en faire un objet à percevoir.



*Illustration 2: Boucle sensorimotrice*

Pour cela, nous nous plaçons dans le contexte d'une théorie sensorimotrice de la perception dans laquelle l'action occupe une place prépondérante. La perception n'est pas passive. Il ne suffit pas simplement d'ouvrir les yeux pour percevoir ce qui se trouve en face de nous. Au contraire, percevoir est une activité basée sur un aller-retour entre un sujet et le monde qui l'entoure. Le sujet agit dans le monde sur le milieu et il obtient en retour des sensations. Lorsque que les actions conduisent toujours aux mêmes sensations, la boucle se stabilise en un invariant sensorimoteur (une loi de contingence sensorimotrice). Cette régularité dans la boucle est la condition à

laquelle il est possible de percevoir [O'Regan 2001]. De ce fait, percevoir le web revient à trouver des régularités dans ce milieu par un jeu d'actions et de sensations. Dans la mesure où il n'est pas possible d'agir directement sur le milieu, le sujet percevant est couplé avec un dispositif qui donne un pouvoir d'action spécifique. Dans ce cas, les actions sur le milieu sont effectuées via le dispositif qui devient une prothèse. Dans le cas d'un système complexe comme le web, le dispositif proposé est une interface interactive qui propose un jeu d'actions desquelles naîtra une perception du système.



*Illustration 3: Deux objectifs dans la construction des visualisations*

Afin d'être la plus efficace possible, l'interface fait appel à la visualisation d'information (infoviz en anglais pour information visualization) définie comme l'utilisation de représentations visuelles, interactives, et sur ordinateur, de données abstraites pour amplifier la cognition [Schneiderman 1999]. Le caractère synoptique de la perception visuelle couplé aux possibilités du numérique en terme de manipulation et de présentation dynamique font de ce couplage un outil capable de faire naître des connaissances qu'il serait très difficile, sinon impossible, d'obtenir [Ware 2004]. Les visualisations peuvent prendre différentes formes : diagrammes, tableaux, cartes, graphes, etc., et peuvent être bidimensionnelles ou tridimensionnelles suivant le but que l'on recherche. Ici deux types de visualisations sont envisagées dont l'objectif final n'est pas le même : soit l'on souhaite disposer d'une visualisation ouverte pour découvrir des prises sur l'organisation globale, soit l'on souhaite utiliser et visualiser le résultat sur l'organisation d'un descripteur particulier et dans ce cas on utilisera un second type de visualisation.

Ces dernières sont construites à partir de données de base, sur lesquelles on applique des filtres ou des algorithmes de traitements, que l'on visualise pour en prendre connaissance. Dans ce cas, il s'agit de transmettre une connaissance particulière sur le système et la visualisation doit être contrainte perceptivement pour qu'il ne soit pas possible de tirer des conclusions erronées. Un diagramme en deux dimensions est un exemple de visualisation contrainte. S'il est bien fait, il n'est pas possible d'en comprendre autre chose que ce qu'il présente. Ce n'est pas le cas des graphes ou des visualisations en 3 dimensions qui peuvent être mal interprétés. Une forme peut par exemple apparaître avec un angle de vue particulier mais il se peut qu'elle ne soit que le fruit d'un jeu de perspective. Un utilisateur incapable de changer le point de vue et qui ne ferait pas attention pourra par ce biais se construire de fausses connaissances et c'est ce qu'il faut à tout prix éviter. L'interactivité est donc réduite de même que la 3 dimension et les visualisations ouvertes car le concepteur n'est pas capable de penser à l'avance toutes les connaissances qui pourraient naître de telle ou telle manipulation et cela perdrait alors son intérêt. Pour réduire au maximum les chances de mauvaises interprétations, ces visualisations sont construites suivant les règles de base du design graphique [Bertin 1967][Tufté 1993]. Bertin a proposé le découpage de chaque problème en composantes typés (sélectives, ordonnées, quantitatives) et fait

correspondre à ces composantes des variables visuelles (taille, valeur, grain, couleur, orientation et forme) de même type. Des scores numériques par exemple sont une composante quantitative et ne peuvent pas être représentés par des formes ou des couleurs car ces variables ne sont pas mesurables. On ne peut pas les classer en les voyant : le bleu ou le rouge ou encore un rond et une étoile ne possède pas une différence pouvant être mesurer au contraire de la taille qui elle est mesurable. Tufte a aussi proposé des règles pratiques comme le pourcentage d'encre vraiment utilisée pour représenter une information en regard de l'encre totale utilisée dans la page. Il prescrit aussi l'utilisation de moiré (vibrations créées par une alternance de bandes noirs et blanches) qui trouble le lecteur. Fort de ces recommandations les visualisations se veulent efficaces et les plus contraintes possible.

Le second type de visualisations est très important pour explorer les systèmes complexes : il s'agit des visualisations comme espace d'actions. Dans ce cas, la visualisation est utilisée pour changer l'algorithme utilisé ou bien pour modifier ses paramètres ce qui a pour effet de changer les données visualisées en retour. Les actions sur ces visualisations donnent lieu à des sensations différentes. Si des régularités sont détectées visuellement, alors naît une nouvelle connaissance, autrement dit une nouvelle hypothèse sur l'organisation que l'on pourra tester de manière spécifique. Cette dernière provient directement de la manipulation. Ces visualisations sont donc ouvertes, interactives et exploratoires pour permettre aux utilisateurs de se construire des espaces d'actions et donc de perception. Ce type de visualisations est construit suivant les règles de l'infviz [Schneiderman 1999][Munzner 2000] et elles sont principalement composées de graphes mais pas seulement.

L'interface finale est un mélange de ces deux types de visualisations et est elle-même une visualisation interactive synoptique de l'ensemble du processus. L'utilisateur est alors capable d'appréhender facilement l'enchaînement des traitements qu'il met en oeuvre. Les traces de ses actions sont directement accessibles et il est plongé au coeur de son processus de recherche, en immersion, ce qui limite les risques d'erreurs liés à la méconnaissance des actions précédentes, fréquentes lorsque l'on enchaîne beaucoup de traitements. Le processus global n'est pas fait de traitements automatiques et l'utilisateur intervient à tous les niveaux en manipulant l'interface, les traitements et les visualisations. Le dispositif est centré sur l'utilisateur et l'automation n'a de sens qu'une fois l'enchaînement des filtres bien stabilisé. Des traces ont été ajoutées à cette trace globale pour permettre à l'utilisateur mais aussi aux chercheurs s'intéressant à la découverte de connaissances de rejouer le film de l'exploration (l'enchaînement des algorithmes et visualisations) afin de comprendre comment et pourquoi les connaissances sont nées. La trace globale est bien entendu archivée ainsi que toutes les actions sur les filtres (paramètres) mais ce n'est pas tout. Sont aussi conservés l'ensemble des déplacements et des actions sur les visualisations, de sorte qu'un observateur puisse comprendre comment l'utilisateur a pu faire naître des connaissances qui proviennent des manipulations sur les espaces d'actions. L'interface finale est donc un outil utilisé à la fois par les explorateurs experts et par les chercheurs pour comprendre et pouvoir reconstruire à des fins de

capitalisation, le processus global d'appréhension du système complexe.

Fort des ses ancrages théoriques, nous avons construit une méthode d'exploration du web en trois temps : rendre disponible les données, les explorer, et enfin manifester leur essence aux utilisateurs finaux.

## Cliché du web

La première étape en amont de l'exploration consiste à prendre un cliché du web qui sera ensuite rendu disponible via une base de données. Ce cliché consiste à indexer l'ensemble des pages hypertextuelles du web, ce qui est bien entendu impossible compte tenu du nombre de pages qu'il contient. Google déclare indexer à ce jour environ 8 milliards de pages et Yahoo plus de 20 milliards (information disponible sur leurs sites). D'autres études font état de 11,5 milliards de sites indexables directement [Gulli 2005] et le chiffre peut aller jusqu'à 550 milliards si l'on considère les pages générées dynamiquement (appelées le "deep web") [Berkman 2001]. Les moteurs de recherche n'indexent donc selon toute vraisemblance pas tout le web et leur taux de recouvrement diffère suivant l'estimation de la taille du web que l'on choisit. La solution est alors d'indexer un sous-graphe du web et de travailler sur ce sous-graphe uniquement, en le complétant au besoin. Pour ce faire on utilise la démarche suivante : on choisit des pages comme points d'entrées, on envoie des robots à partir de ces pages qui vont suivre tous les liens hypertextes et indexer tous les contenus dans une base de données [Druegeon 2005]. Le problème est encore une fois le même que les mésopotamiens. A quelle profondeur en terme de nombre de clics faut-il aller, où doit-on s'arrêter ? Si l'on estime que le diamètre du web est de 19 clics [Albert 1999] alors une profondeur 19 reviendrait à indexer tout le web. On observe pourtant empiriquement qu'au delà de 6 clics, les robots ramènent énormément de documents, tant et si bien que l'on limite généralement la profondeur à 3 clics (en termes de sites et non pas de pages). D'autres problèmes viennent se superposer à cette question de taille. Les pages sont organisées en une unité sémantique et sémiotique que l'on appelle site Internet. Une solution pour réduire le nombre de documents à indexer est alors de ne considérer que les sites et non plus les pages. On peut conserver les pages mais la représentation de la connectivité est limitée aux sites. Du coup le stockage des données est plus réduit. On peut également choisir de ne considérer que les liens et un ensemble de mots clés pour éviter de stocker toute les pages en entier. De plus le web étant très dynamique, les changements dans les pages, d'un crawl à un autre, sont consignés et indexés aussi. Dans une mesure générale, les parcours des robots (crawl) ont pour but d'indexer tout le web sur un sujet particulier qui sera ensuite exploré. Pour cela on demande à des experts de ce domaine de donner leurs points d'entrées (des pages pertinentes). Ensuite les robots partent de ces pages et suivent les liens. Si l'on s'aperçoit que le contenu sémantique est trop éloigné du sujet de départ, le robot peut ne plus poursuivre (principe du focused crawling [Chakrabarti 1999]) dans cette voie et se recentrer sur d'autres pages.

Pour être certains d'avoir capté toutes les ressources intéressantes, le crawl est généralement lancé plusieurs fois et l'on détecte ainsi si des régularités apparaissent.

Le processus de détection des régularités, décrit dans la section précédente, permet de stabiliser le sous-graphe du web extrait. Toutes les informations relatives au processus d'indexation des pages sont bien entendues conservées et synthétisées/manifestées en direction de l'utilisateur pour qu'il sache à tout moment sur quelles bases il travaille. Cette étape est encore largement expérimentale et basée sur le travail d'un utilisateur qui seul décide quand le cliché est satisfaisant, fort des indices statistiques et des visualisations dont il dispose. Lors d'une étude sur le débat sur le traité constitutionnel européen en ligne, nous avons par exemple crawler plusieurs fois. À chaque tentative, on regarde si dans la dernière profondeur explorée, on trouve encore des sites intéressants (à l'aide d'outils statistiques sur les mots clés mais aussi en allant visiter les pages dans un navigateur classique). Si c'est le cas alors on relance le robot dans cette direction en lui donnant en entrées les sites qui se trouvaient à la dernière profondeur. On relance alors un crawl en profondeur 1 ou 2. On tente ainsi de trouver une frontière à partir de laquelle les sites ne sont plus en rapport avec le sujet. Ce n'est que lorsque l'on pense avoir délimité cette limite pour tous les sites que le cliché est jugé satisfaisant. C'est à cette condition que le travail d'exploration peut commencer.

## Exploration

Les sous-graphes du web obtenus à l'étape précédente doivent maintenant faire l'objet d'une analyse pour en déduire l'organisation. Dans la mesure où cette étape résulte à chaque instant des choix de l'utilisateur, c'est bien une exploration qui procède d'étapes qui vont d'hypothèses en mises à l'épreuve qui infirment ou confirment une tendance. Chacune de ces étapes permet de passer à la suivante par une rupture en transformant la « matière » en « forme », ainsi que l'a décrit [Latour 1996] pour l'activité scientifique en général. Ici la matière est constituée des données de l'étape et leur forme est la visualisation qui leur est associée, de manière plus ou moins forte suivant le type de visualisation choisie.

Chaque rupture est un nouveau traitement qui représente une nouvelle hypothèse et qui occasionne de nouvelles données qui prennent formes dans une nouvelle visualisation.

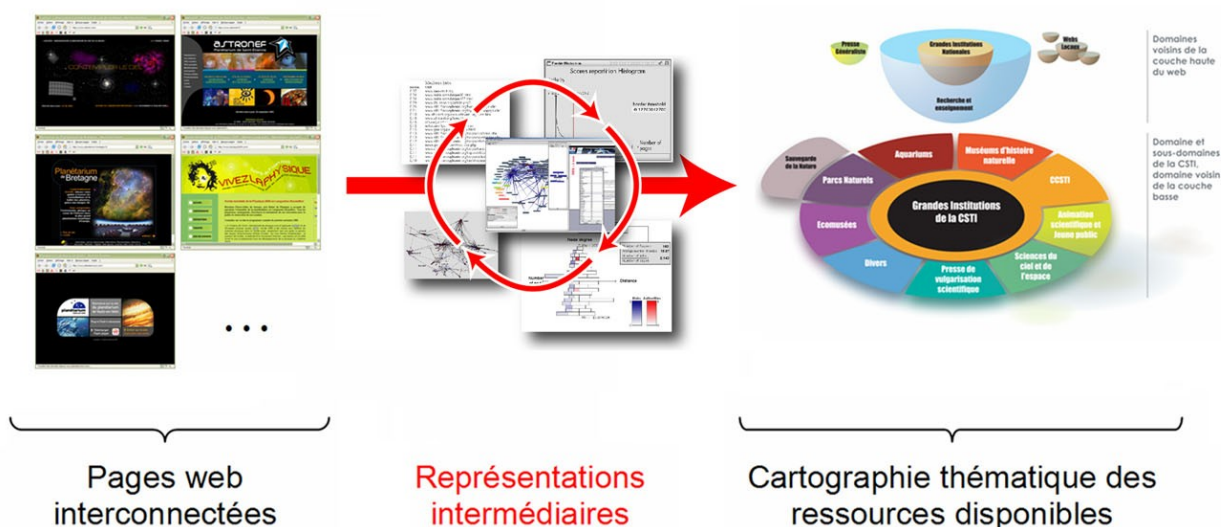


Illustration 4: Du web immersif au web synoptique

## *la visualisation en graphes*

La taille du corpus (le sous-graphe d'Internet) que nous indexons change en fonction du sujet. Dans le cas de l'étude sur le traité constitutionnel européen, le corpus était de 2,5 millions de pages et 12 000 sites. Une fois indexé, ce sous-graphe d'Internet est mis en forme grâce à un logiciel d'affichage de graphes, dont la principale fonction est d'organiser spatialement le graphe sur l'écran de l'ordinateur pour le rendre lisible. Cette fonction appelée « force vector » applique des lois physiques à tous les nœuds du graphe : les nœuds se repoussent mais les liens les attirent. Ainsi manifesté dans un espace pseudo-physique, le sous-graphe du web devient manipulable de façon assez semblable à un objet matériel. La suppression d'un nœud entraîne la réorganisation de tout le graphe (on peut se représenter un tel graphe actif comme une sorte de sommier à ressorts tendu dans le vide).

Ce graphe actif est manipulé pour faire émerger des motifs significants. Les opérations effectuées par l'utilisateur du dispositif sont à la fois algorithmiques (puisqu'elles mettent en jeu des calculs sur les données) et graphiques (puisque les calculs sont toujours manifestés visuellement). Les scores résultant des différents algorithmes peuvent être manifestés sur des graphes en particulier par l'emplacement des nœuds, leur couleur, leur taille ou d'autres dispositifs plus sophistiqués. Détaillons les opérations les plus courantes. Premièrement, les graphes doivent être « nettoyés ». Lorsqu'il s'agit de dégager une structure, les sites qui ne sont connectés qu'à un seul autre site ne sont pas pertinents : ils sont donc effacés, rendant le graphe plus lisible. Deuxièmement, les nœuds restants ne sont pas répartis de façon homogène. Certains se chevauchent, d'autres s'illustrent malgré leur manque d'importance. Il arrive souvent que l'on trouve plusieurs composantes connexes et des zones de regroupement apparaissent. On s'intéresse alors à chaque composante connexe individuellement puis on sépare visuellement les sites qui se regroupent pour les traiter à part. Pour ce faire on utilise des indicateurs comme le score d'autorité (HITS [Kleinberg 1998], voir plus bas) qui représente l'importance d'un site dans la structure hyperliée : tous les sites ayant un faible score sont éliminés et les sites restants sont classés à la main par thématiques. Cette opération permet d'appréhender le ou les « coeurs » d'un domaine. Troisièmement, les sites qui ne font pas partie du ou des coeurs doivent être également explorés. Plusieurs techniques sont utilisées, citons-en quelques unes : l'exploration du voisinage de sites du coeur de même thématique, l'exploration du graphes des liens bidirectionnels, l'exploration du graphe des sites de faible score d'autorité, etc. Dans tous les cas, il s'agit de trouver à la fois des regroupements thématiques de sites et les thématiques qui permettent au mieux de faire des regroupements. Enfin ce sont les relations de voisinage entre ces regroupements thématiques qui sont étudiées. A la fin de cette étape, le corpus initial du traité constitutionnel européen ne compte plus que 295 sites. L'étape suivante consiste à manifester ce corpus aux utilisateurs finaux.

Le cœur du processus d'analyse et d'interprétation du web se trouve donc au niveau de l'exploration du graphe. Il y a plusieurs raisons à cela et la première est que le graphe actif est une présentation à la fois synoptique et analytique du web. Il est possible de distinguer des formes générales mais il faut pour cela manipuler le graphe et repérer les régularités. La seconde raison est que par nature, le graphe permet une

double mise en présence du web par l'espace graphique du graphe et par la possibilité d'obtenir les pages Internet via le navigateur classique, auxquelles l'accès est facilité par l'équivalence page-nœud. Le graphe actif est susceptible d'être manipulé graphiquement pour faire apparaître diverses caractéristiques des pages.

Au final, la polyvalence du graphe le rend particulièrement efficace pour permettre une activité d'expertise basée sur la mise en co-présence de l'expert et de son objet.

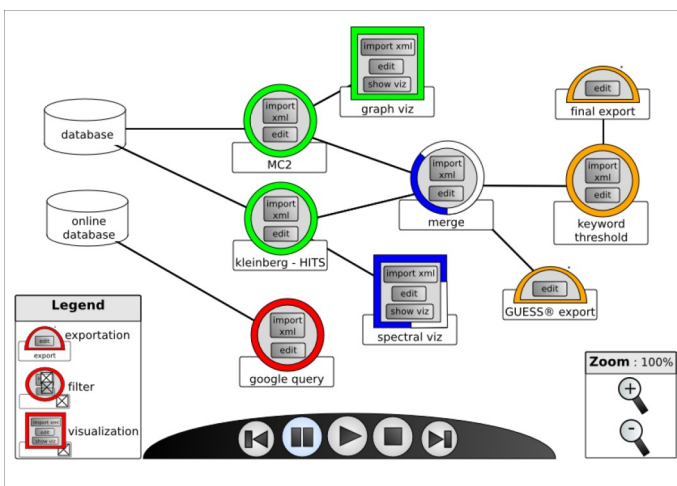
### *Concevoir un logiciel d'exploration*

La visualisation en graphe est donc un outil efficace en terme de découverte de connaissances et tout particulièrement pour la recherche de motifs. De manière plus précise, le graphe est utilisé pour rendre perceptibles des corrélations entre la topologie, la sémantique et la sémiotique des pages ou sites web. Ces types de descripteurs sont combinés et manifestés ensemble dans un même espace. Dans la mesure où nous avons à faire à un système dont nous ne connaissons pas l'organisation entre les données, la tâche principale s'oriente vers la découverte de descripteurs permettant d'expliquer les relations entre ces données. Les graphes sont des objets ouverts et manipulables sur lesquels les actions de changement de points de vue ou de transformation de la présentation peuvent faire ressortir des régularités. Autrement dit la boucle sensorimotrice sur la visualisation en graphe fait naître une perception qui est précisément une perception de la corrélation inscrite au départ. Cette perception est ensuite traduite en un descripteur du système qui représente une hypothèse sur l'organisation. Reste à la vérifier et c'est là le rôle de la première catégorie de visualisations qui visent à présenter le résultat d'un traitement de manière synthétique.

Dans le processus général, nous avons donc des données à notre disposition et des hypothèses de départ sur leur organisation. C'est le cas en topologie par exemple où [Kleinberg 1998] a proposé de considérer que les sites peuvent être classés en fonction de leur nombre de liens entrants et sortants (un portail ou hub aura beaucoup de liens sortants et une autorité beaucoup de liens entrants) et que cela fait sens dans l'organisation du web. Le pagerank de google est un autre exemple d'hypothèse qui considère les liens entrants dans un site comme un indice de son intérêt et attribue le score en fonction de ce critère ne le propageant. Un site voit son pagerank augmenter si les sites qui le cite ont eux aussi un pagerank élevé. Il faut noter que le pagerank n'est depuis quelques années plus uniquement un critère topologique mais aussi sémantique qui tient compte également des mots clés dans la page et dans les balises méta suivant un principe que l'on ne peut qu'inférer. Google ne dévoile en effet plus ses algorithmes pour éviter les tricheurs qui réduisent la pertinence des résultats ou bien que d'autres moteurs utilisent le même descripteur. Fort des hypothèses de départ, l'exploration consiste à mettre en place une double stratégie.

Il faut dans un premier temps se donner les moyens de vérifier la pertinence de ces descripteurs ou hypothèses. C'est le rôle des visualisations contraintes. Elles permettent de vérifier des traitements construits sur la base du descripteur de la manière la plus juste qui soit. Si l'on souhaite vérifier que la densité de connectivité sur le web suit bien une loi de puissance que ce soit entre les sites ou au sein d'un site

[Barabasi 2002] par exemple, une visualisation en graphe ne sera pas très pertinente, contrairement à une visualisation en histogramme où l'on visualise plus sûrement les évolutions. La liberté de perception face à ce dernier type de présentation est très faible. Ensuite, si les descripteurs existants ne suffisent pas (et le système n'est plus complexe si ce n'est pas le cas) alors il faut trouver de nouveaux descripteurs. On fait ici appel au second type de visualisations. La dimension exploratoire décrite ci-dessus permet de découvrir de nouvelles prises, de nouveaux descripteurs qui à leur tour vont faire l'objet d'un test en visualisation contrainte. Il faut préciser ici que les tout premiers moments de l'analyse d'un sous-graphe du web sont purement exploratoires et non pas analytiques. Il est nécessaire de se familiariser avec le graphe, de se forger des points de repère pour s'orienter dans le travail, d'aller voir les sites principaux, d'avoir un premier rapport avec l'objet avant de tenter de l'expliquer. Tout participe de la saisie de l'objet et doit la faciliter.



*Illustration 5: Projet d'interface de manipulation du processus général (Ngiyari) en cours de réalisation*

La boucle entre découverte de descripteurs, leur utilisation et leur validation, s'arrête lorsque les descripteurs trouvés permettent d'expliquer de manière satisfaisante l'organisation du sous-graphe du web que nous avons indexé. Il s'agit là encore d'une application de la théorie sensorimotrice de la perception à un échelle plus large. Les actions sont les différents descripteurs mis en oeuvre et les sensations sont l'allure du jeu de données ainsi constitué (les visualisations). La perception résultante est une appréhension de l'ensemble du processus autant que du sous-graphe indexé lui-même et donc d'une partie du web. Afin de favoriser cette exploration de second ordre, l'interface de manipulation des traitements est elle-même libre et favorise la perception de l'ensemble de la chaîne. La trace est disponible tout au long du processus et elle est en plus de cela visible et donc perceptible. Toutes les visualisations sont, de plus, présentes en même temps. La conséquence de cela est que l'on dispose à tout moment de différentes vues sur un phénomène unique. Ces présentations vécues ensemble comme autant de points de vues sur une même organisation réelle du web enrichissent la préhension en la facilitant.

## Conclusion

Une fois le sous-graphe appréhendé, son organisation est mise à jour et l'on dispose d'un corpus de documents dont on connaît la structure. Vient alors la dernière étape qui consiste à rendre compte de cette structure. Dans le cas du web, cela signifie être en mesure d'expliquer le sous-graphe étudié. Ici encore tout dépend du besoin. Certains souhaiteront disposer d'indicateurs statistiques sur le sous-graphe et d'autres voudront pouvoir y naviguer mais de manière efficace et sans se perdre. Dans la mesure où l'explorateur expert sait quels traitements il a effectué pour parvenir à

son résultat, il faut pour qu'elle soit comprise que son explication reflète ces traitements. C'est le cas avec la carte présentée dans l'illustration 6. L'exploration précédant la carte a été effectuée à partir du descripteur de Kleinberg avec les portails et les autorités, que l'on retrouve sur la carte car ce descripteur sert de métrique pour classer les sites du centre à la périphérie. De la

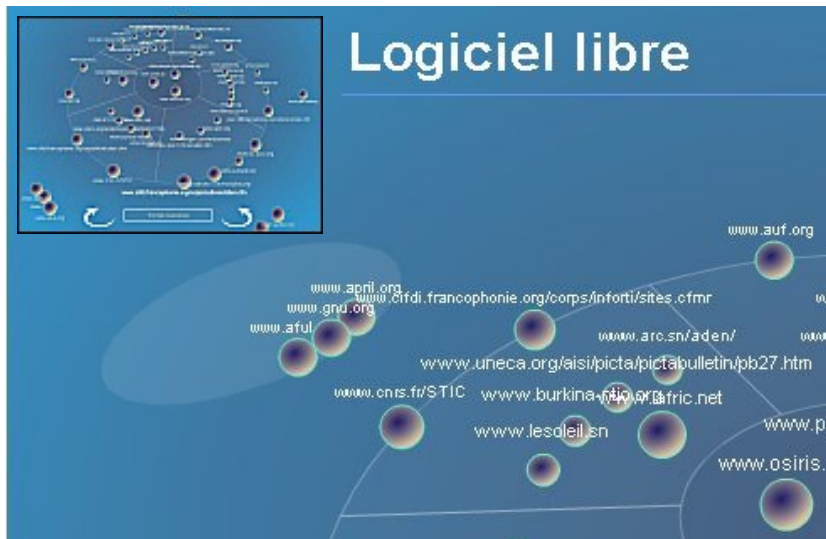


Illustration 6: détail de la carte de la coopération nord-sud TICE sur internet, sous-domaine du logiciel libre.

même façon les sites ont été classés sémantiquement en catégories et l'on retrouve ces catégories comme des quartiers sur la carte. Cette visualisation est également interactive. Elle permet d'afficher différentes informations quant à la nature des sites et de les charger dans le navigateur. On se trouve donc face à une boucle où l'on est parti du web, on l'a isolé, exploré et analysé pour finalement y retourner mais cette fois-ci avec un guide. Ce dernier est d'autant plus pertinent qu'il présente la structure de la portion du web étudiée.

Au final, pour celui qui expertise un sous-graphe du web, la problématique n'est pas d'appliquer des traitements automatisés mais de s'engager dans les « plis » du web et de constituer des « prises » qui puissent être transmises. Il s'agit d'explorer le web en l'appréhendant par divers outils, et d'identifier des traces intéressantes. Ces traces sont tout d'abord diffuses, et par les différentes étapes que l'expert effectue, il cherche à reporter ces traces de représentation en représentation jusqu'à produire des vues synoptiques du domaine et des « zooms » sur des saillances pertinentes. Au cours de ce processus, le web perd en multiplicité, en flexibilité, en complexité, mais gagne en universalité et en lisibilité.

Cette méthode permet donc à partir du web d'en tirer l'essence via des dispositifs adaptés qui donnent une large part aux visualisations. L'usage à bon escient de ces dernières permet de se saisir de l'objet étudié et d'en faire un objet rationnel qui n'échappe pas à notre prise. Les visualisations sont à la fois des outils de capitalisation et de production de traces qui permettent aux utilisateurs de s'en saisir et de suivre des fils, des intuitions, sans se perdre. Une première version de l'interface existe déjà dans le logiciel TARENTE [Ghitalla 2004] et nous sommes en train de développer une seconde version baptisé FEW<sup>3</sup> (Framework for Exploring World Wide Web) en partenariat avec l'INA.

## Bibliographie

Albert, B., Barabasi, A.-L., Jeong, H., *Diameter of the World Wide Web*. Nature, 401(6749):130--131, September 1999.

- Barabasi, A. L.**, *Linked : The New Science of Networks*, Perseus Publishing, 2002.
- Bottero, J.**, *Mésopotamie, L'écriture, La raison et les dieux*, Gallimard, 1987.
- Bergman, M. K.**, *The deep web: Surfacing hidden value*, The Journal of Electronic Publishing, 2001
- Bertin, J.**, *Sémiologie Graphique. Les diagrammes, les réseaux, les cartes*. Paris-La Haye : Mouton, Gauthier-villars, première édition, 1967.
- Chakrabarti, S., van der Berg, M., & Dom, B.**, *Focused crawling: a new approach to topic-specific Web resource discovery*. In Proceedings of 8th International World Wide Web Conference (WWW8), 1999
- Drugeon, T.**, A technical approach for the french web legal deposit. IWAW 2005.
- Ghitalla, F., Diemert, E., Maussang, C. , Pfaender, F.**, *TARENTE : an Experimental Tool for Extracting and Exploring Web Aggregates*, ICTTA'04, Damas, Syrie, 2004.
- Ghitalla, F., Lenay, C.**, *Les territoires de l'information. Navigation et construction des espaces de compréhension sur le web*, dans *La navigation*, Les cahiers du numérique, Hermès Edition, Paris, 2003
- Gulli, A., Signorini, A.**, *The Indexable Web is More than 11.5 Billion Pages*, in Proc of WWW 2005 Conference, ACM, Chiba, Japan, 2005.
- Kleinberg, J.**, *Authoritative Sources in a Hyperlinked Environment*, in Proc. of the ACM-SIAM Symposium on Discret Algorithms, ACM Press, 1998.
- Latour, B.**, *Le pédofil de Boavista, montage photo-philosophique*, in *Petites leçons de sociologie des sciences*, Coll. "Point", Seuil, p171-225, 1996.
- Lenay, C.**, *Constitution de l'espace et immersion*, Arob@se, [www.univ-rouen.fr/arobase](http://www.univ-rouen.fr/arobase), volume 1, pp. 85-93, 2005.
- Lorenz, H.**, *Predictability: Does the Flap of a Butterfly's Wings in Brazil set off a Tornado in Texas?*, meeting of the American Association for the Advancement of Science in Washington, D.C., 1972.
- Munzner, T.**, *Interactive Visualization of Large Graphs And Networks*, Ph.D. Dissertation, Stanford University, June 2000
- O'Reagan, K., Noë, A.**, *A Sensorimotor account of vision and Visual Consciousness*, Behavioral and Brain Sciences, 24 (5), 2001.
- Shneiderman, B., Card, S.-K., MacKinlay, J.-D.**, *Readings in Information Visualization, Using Vision to Think*, Morgan-Kaufmann Publishers, New-York, 1999.
- Tufte, E.** *Visual Display of Quantitative Information*. Graphic Press, Cheshire, connecticut, seconde edition, 1993.
- Ware, C.**, *Information visualization: Perception for design.*, Morgan-Kaufmann Publishers, San Francisco, 1999.