Chroniques du Web

Expéditions scientifiques dans l'univers des données numériques et des réseaux
- Carnets cartographiques -

Prologue

Chroniques du web (première partie)

Etudes Cartographiques (seconde partie)

Prologue

Ce livre n'est pas un ouvrage scientifique, du moins au sens classique. Il s'agit plutôt du récit d'une quête curieuse animée par une question qui m'obsède: le web, ce vaste système distribué de documents et de contenus, peut-il avoir une *forme*? Peut-on en décrire l'organisation, les propriétés, la *géographie*? Le web, la couche la plus abstraite de l'Internet, peut-il être décrit comme un *espace* ou un *objet*, un *territoire* ou un *processus*?

Dès mon arrivée à l'Université de Technologie de Compiègne en 1997 comme chercheur en Sciences de l'Information, j'ai intégré le département *Technologie et Sciences de l'Homme*, ce genre de département qu'affectionnent les écoles d'ingénieurs censé apporter aux étudiants ce «supplément d'âme» nécessaire à la compréhension des enjeux politiques, économiques ou sociaux des technologies. J'ai donc commencé par enseigner la *linguistique et la philosophie du langage* puis, dans un second temps, les *usages des technologies numériques de l'information*. Tout aurait dû me conduire à développer des recherches depuis cette *périphérie de la technologie* où l'on examine ses usages sociaux, les opérations cognitives qu'elle supporte, les interfaces par lesquelles elle nous apparaît, l'histoire et l'épistémologie qui en décrit les périodes ou l'anthropologie qui en étudie le différents rôles. Un réseau ouvert de documents numériques comme le web aurait donc pu constituer pour moi un excellent terrain d'étude du point de vue des sciences humaines et sociales: un terrain d'observation encore vierge à l'époque que j'aurais décrit avec ce regard critique nourri d'érudition qui fait la spécificité des SHS. Oui, mais voilà: au début des années 2000, lors que je commençais à enseigner en école d'ingénieurs, le web entamait sa formidable expansion et posait des questions massives auxquelles il allait falloir se confronter et auxquelles n'étaient pas préparées les sciences humaines et sociales, ni même peut-être nos universités: ses masses, son caractère distribué et son évolution temporelle, ses technologies, le modèle d'innovation qui sous-tend une telle dynamique de développement...et les raisons des premiers succès des géants californiens de l'information qui dominent encore aujourd'hui le monde des moteurs de recherche et des services. Faute d'ingénierie spécifique et de programmes technologiques ambitieux en France comme en Europe, les sciences humaines n'ont pas su inventer les instruments originaux et l'assise expérimentale qui aurait permis de nourrir de leurs spécificités toutes les innovations qui sont sorties de ce vaste univers de *data* et de services. L'idée de *produire* la technologie plutôt que de la *décrire* s'imposa donc clairement à moi et, si le web posait autant de questions, autant tenter d'y répondre *de l'intérieur*.

Cet univers m'intriguait d'autant plus que tout le monde reconnaissait l'importance de ce réseau mais s'évertuait à le représenter sous la forme d'un nuage énigmatique. Au printemps 1999, j'entrepris donc de modifier la forme de mon enseignement et ses objectifs pour me concentrer sur un défi dans lequel les élèves ingénieurs me suivirent naturellement: *dessiner le web*! Tout fut bon pour arriver à un schéma ou un plan intelligible plutôt qu'un nuage indistinct: défricher la littérature scientifique sur le sujet, suivre de près les développements technologiques dans l'univers des premiers moteurs de recherche et, surtout, construire nous-mêmes ces petits programmes qui permettent de naviguer sur le web et de récolter nos propres jeux de données, les *web crawlers*. Petit à petit, je me mis à valider les étudiants autant par le *faire* que par le *dire* ou *l'écrit* et, à ce jeu-là, tout me convenait qui puisse ouvrir l'imagination à cette sorte de *cybergeography* du web: expériences, analyse de données, algorithmes mais aussi plans, schémas et même maquettes! En découvrant les premières réalisations des étudiants, j'acquis la conviction que l'effort exigé pour représenter graphiquement le web constituait une étape précieuse pour développer ensuite les technologies et les services qui seraient adaptés à sa morphologie, à son architecture, à sa géographie documentaire. Une porte ouverte sur l'innovation.

Mon enseignement ressemble depuis lors à un atelier de créativité d'où sont sortis des projets qui ont rythmé ma vie, et nourri aussi quelques convictions. La première est qu'il est souvent bien difficile de trouver une place à l'innovation dans nos université et nos écoles, toujours coincée entre «enseignement» d'un côté et «recherche» de l'autre. En modifiant mon enseignement, je transformai donc aussi mes pratiques de chercheur pour construire cet espace nécessaire à nos nouvelles activités où nous avons emprunté de nombreux chemins et où le hasard eut sa place. La seconde est qu'il ne faut pas limiter l'*innovation* aux cadres de «l'innovation industrielle», comme trop souvent le cas en école d'ingénieurs. Comme le démontrent aujourd'hui les technologies web dans leur transformation du monde, l'innovation peut-être tour à tour, et sans hiérarchie, *industrielle* mais aussi *scientifique*, *sociétale*, *culturelle* ou *politique*. Il faut donc admettre que l'innovation technologique ne se développe efficacement que dans un cadre «d'enseignement» ou de «formation» profondément réinventé. De là, une troisième conviction: former des ingénieurs (et peut-être au delà toute la jeunesse), c'est admettre avec eux une forme *d'ignorance partagée* sur quelques-unes des problématiques fondamentales qui marquent notre temps, comme les problèmes climatiques et environnementaux, la biodiversité en danger, les méfaits de l'économie spéculative ou...les zones d'ombre qui masquent encore quelques-unes des propriétés étonnantes du web comme sa dynamique évolutive. Soyons francs: face à de tels défis théoriques ou sociétaux, et à relever si rapidement si nous ne voulons pas en payer le prix fort, peut-on faire autre chose que de partager notre ignorance avec une jeunesse dont l'appétence pour le monde de demain est sans limite?

*****

La forme de cet ouvrage peut étonner: une série de chapitres rédigés de façon libre et personnelle où le *je* comme le *nous* alternent souvent indistinctement, pour marquer la dimension collective d'une série d'aventures expérimentales. Pour parler vrai, ni le «je», ni le «nous», et encore moins le genre de recherche technologique dont il est question dans cet ouvrage n'ont vraiment de place dans la recherche scientifique française, cloisonnée par ses disciplines, académique dans ses procédures, figée dans son organisation, conservatrice dans ses formes d'expression. C'est pourquoi on ne trouvera guère dans les revues officielles *nos* contributions, ces fameux «articles» dont le nombre et la supposée qualité sont censées m'évaluer à titre individuel (tout à fait réglementairement puisque je suis fonctionnaire) et constituer des indices fiables de l'*excellence scientifique*. Outre le caractère absurde de l'expression en science, les textes officiels nous enjoignent de «rayonner» à travers les canaux restreints de la publication académique, sans tenir compte, à aucun moment, de toutes ces externalités par lesquelles nous diffusons les démarches, nous renouvelons les méthodes, nous inventons les instruments de demain, nous créons de l'activité et des emplois. *Rayonne-je*? La question m'apparaît aussi ridicule que sa prononciation. Et puis rayonner dans quel type d'espace? Je ne sais toujours pas aujourd'hui à quel type exact de recherche nous nous sommes livrés pendant des années, et pas plus aujourd'hui. Notre bricolage inventif, basé sur une ignorance partagée à l'égard d'un réseau qui semble encore échapper à notre compréhension, se situe pourtant quelque-part entre «recherche» et «enseignement» dans nos universités, entre Sciences Humaines et sciences informatiques mais sans que je n'ai jamais pu rattacher notre projet aux unes ou aux autres. Un *quelque-part* qui n'existe, à vrai dire, ni dans les programmes de recherche européens qui petit à petit nous sont imposés (le fameux *H2020*-Horizon 2020 de la recherche européenne), ni dans les programmes régionaux et encore moins

dans les programmes redondants de l'Agence Nationale de la Recherche (A.N.R.) dont les financements sont systématiquement accordés aux mêmes équipes depuis toujours.

Le web fut donc salvateur à bien des égards. Dès le départ de nos aventures, dans les années 2002-2003, il a constitué notre objet d'étude favori mais aussi un excellent vecteur de diffusion de nos trouvailles techniques et méthodologiques. Je ne compte plus aujourd'hui le nombre de fois où des visualisations de graphes issues de *Gephi* (notre application favorite) ont été publiées dans *Sciences, Nature*, *National Geographic*, *Forbes* ou le *Washington Post* à propos de *Social Network Analysis* mais aussi de génétique, de biologie, de sciences de l'environnement, d'économie ou de *data sciences*. Je n'ai jamais non plus identifié précisément la façon dont nous avons pénétré la recherche américaine à Stanford, au M.I.T. ou chez *Facebook* ou *Google* mais le réseau a joué un rôle fondamental ici aussi. Je ne connais seulement, ou simplement me rappelle, que les destins personnels de ceux qui ont contribué à dessiner ce qui ressemble à un sorte d'écosystème dont le web est l'ossature: les fondateurs de *Linkfluence*, le Consortium *Gephi*, ceux partis contribuer à la RetD de *LinkedIn* ou d'autres entreprises en Californie, la création de *Linkurious*, la librairie *SigmaJS*, certains outils du Médialab de Sciences-Po comme le *navicrawler*, la naissance récente de l'*Atelier Iceberg*. Nous avons donc étudié mais aussi nourri de nos instruments l'écosystème du web. Ce livre en est issu, presque directement puisqu'une partie des informations contenues dans cet ouvrage sont extraites de mon blog, *L'Atelier de Cartographie* ([1]). Ouvert en 2011, le blog a accueilli depuis lors plus de 42.000 visiteurs uniques et plus de 140.000 visites. Je suis loin, évidemment, du million de téléchargement de *Gephi* rien qu'en 2014! Accumulés chronologiquement, mes posts ont déjà tout d'un récit qu'il s'agissait de mettre en forme à travers cet ouvrage.

L'*Atelier de Cartographie* archive aussi les quelques documents que j'ai produits pour décrire les pistes possibles de développement d'instruments inédits et de méthodes nouvelles en ingénierie de l'information; ils étaient essentiellement programmatiques et tournés vers les applications possibles de nos travaux. Ce type de document et de «posture» comme on dit n'a pas sa place dans la littérature scientifique officielle, y compris en école d'ingénieurs mais peut s'avérer précieux comme source d'inspiration. C'est pourquoi l'idée d'écrire cet ouvrage m'est venue dès 2008 lors de la publication de l'un de mes rares articles dans une revue scientifique, précisément *Communication et Langages* n°158 de décembre de la même année. Mon article était intitulé «L'Atelier de Cartographie» et consacré à la *Toile Européenne*, une étude cartographique des sites web consacrés à l'Europe réalisée par une jeune *start-up* de l'époque, *Linkfluence* qui travaillait en partenariat avec le Centre d'Information sur l'Europe. Le responsable de la revue publie dans le même numéro un article consacré à l'écriture scientifique, à sa nature et à ses règles, pour signifier combien le mien n'entrait pas dans les cadres convenus de l'exercice. Comme l'écrit le directeur de publication dans ce numéro, mon article a fait «l'objet d'échanges animés dans le Comité de lecture». Et il poursuit: «Deux types de questions ont avant tout cristallisé l'attention au cours de ces échanges. Le premier a porté sur la pratique d'écriture des textes scientifiques; c'est une interrogation récurrente dans l'équipe de *Communication et Langages* et plus généralement dans les publications en sciences humaines, l'écriture de certains articles étant parfois jugée trop subjective voire trop littéraire par certains lecteurs. Le second type de questions a quant à lui porté sur la posture énonciative des textes scientifiques. A l'origine de ces remarques, l'article sur l'*Atelier cartographique* consacré aux pratiques et aux enjeux de la «cartographie» thématique de documents web». Il fallait donc trouver d'autres moyens de diffusion de ce que je pense être les sciences *de* l'ingénieur, et non pas des sciences *pour* l'ingénieur. Et, quitte à déroger aux règles du conservatisme académique, autant le faire sous forme d'un récit.

1 Http://ateliercartographie.wordpress.com

*****

Sur ce point, je nous ne faisions que poursuivre le chemin ouvert par des ouvrages légendaires qui ont façonné ce domaine inédit de la *science des réseaux (network sciences)* dans le quel notre projet de carte du web puise son inspiration et une bonne partie de ses méthodes. Plongés dans notre bricolage intellectuel et technique, ils nous ont éclairé à bien des égards en remettant en perspective les nombreuses contributions scientifiques produites à partir de la fin des années 90 en matières de *web sciences* puis de *network sciences*. Nous avons ainsi pu coller, littéralement, à la recherche américaine où se sont hybridées la recherche scientifique et l'ingénierie des *data* donnant naissance à ce que l'on appelle désormais les *data sciences* avec son cortège de déclinaisons comme l'*open data*, les *web sciences*, le *social data mining* ou le *big data*. Ils se sont succédé comme des balises, orientant souvent nos objectifs: *The Laws of the Web* de B.A. Huberman en 2001, *Nexus* de M. Buchanan en 2002, *Linked* de A.-L. Barabasi et *Six Degrees* de D. Watts en 2003, *Sync*. de S. Strogatz en 2004, *The Structure and Dynamics of Networks* l'ouvrage collectif publié en 2006, *Networks* de M.E.J. Newman et Networks, *Crowds, and Markets: Reasoning About a Highly Connected World* de J. Kleinberg et D. Easley en 2010. Nous n'avons jamais eu le temps, ni peut être l'envergure intellectuelle, pour contribuer nous aussi au plan théorique à ces nouvelles sciences basées sur l'ingénierie des données autant que sur les modèles théoriques. Sur la trentaine d'étudiants ou de jeunes ingénieurs qui ont travaillé avec moi, seulement trois se sont inscrits en doctorat, et deux seulement ont aujourd'hui soutenu leurs thèses. A titre de chercheur, je le regrette mais j'ai partagé avec eux cette envie d'aller plus vite pour épouser l'innovation technologique accélérée des californiens, plus loin pour réinventer les méthodes et les instruments associés aux masses de *web data*. Une partie des jeunes ingénieurs qui m'ont accompagnés partagent encore aujourd'hui cette curieuse quête de la forme du web à travers leurs activités professionnelles, chacun à sa façon.

Nous avons donc organisé, dès le début, un modeste atelier artisanal, une sorte de laboratoire collectif hors-murs où nous avons accumulé les projets *ad-hoc* dont chacun a été l'occasion d'enrichir notre boîte à outils méthodologiques et nos instruments. Notre véritable trésor de guerre. Nous n'avons jamais eu d'organisation formelle, hormis l'association *WebAtlas* qui a vécu quelques temps. Ainsi qu'un document d'une page rédigé en décembre 2000 intitulé R.T.G.I. pour *Réseaux, Territoires et Géographie de l'Information*. En revanche, nous avons toujours revendiqué notre rôle dans ce que j'appelle désormais *l'artisanat de haute-technologie*: dans l'univers de l'innovation autour des *web data*, ce sont toutes ces chaînes de traitement de l'information qui, depuis l'extraction des données jusqu'aux interfaces finales, ouvrent sur l'idée de nouveaux services, de nouvelles activités et, je l'espère, de nouveaux types d'emplois ou d'activités. Au cours des années, je me suis spécialisé dans le développement et l'usage intense des instruments cartographiques. N'étant pas suffisamment compétent en mathématiques et en sciences informatiques (je ne suis pas ingénieur!), certains domaines des chaînes de traitement de l'information m'étaient fermés comme le codage d'algorithmes ou la conception d'architectures système. Je me suis donc tourné, petit à petit, vers les instruments de cartographie de l'information ou, pour être plus précis, de *cartographies de graphes relationnels*. Les graphes sont des instruments d'analyse des phénomènes complexes sous l'angle de leurs aspects systèmes (comme peut l'être l'analyse multidimensionnelle). Les aspects mathématiques liés aux matrices de graphes peuvent être d'une grande abstraction (la *théorie des graphes*) mais on peut leur associer aussi des instruments de spatialisation graphique, des sortes de *topographies* ou de

cartographies qui accompagnent les propriétés statistiques alors incarnées visuellement. Le web, à titre de système documentaire ouvert et largement distribué, se prête à merveille à une analyse à base de graphe: les pages web peuvent être désignées comme des *nœuds* reliées entre eux par des *liens* de différentes natures, dont le lien hypertexte qui constitue la colonne vertébrale du réseau. L'infrastructure logistique des moteurs de recherche web fournit, par ailleurs, les bases ou les archives nécessaires aux calculs des propriétés statistiques d'immenses masses d'information. C'est ce passage à l'échelle permis par le web et ses technologies qui, à mon sens, a profondément renouvelé la théorie des graphes et ses instruments mathématiques et techniques.

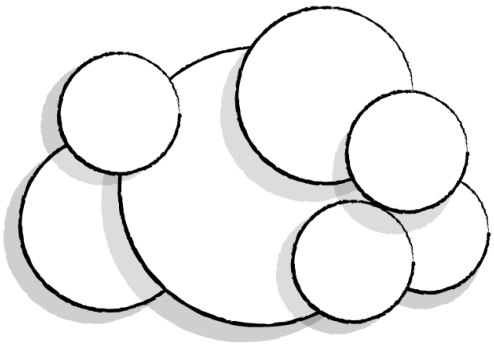
Au début des années 2000, je me suis donc retrouvé dans une situation inédite avec, d'un côté les masses de données web désormais accessibles puisque le réseau est ouvert à l'exploration et, de l'autre, les premières applications de visualisation de graphes dont *Gephi* a désormais marqué l'histoire. Je suis donc devenu une sorte de «photographe du web» pendant plusieurs années, avant d'aborder plus récemment d'autres univers documentaires. Pour moi, ce furent autant d'expéditions dans l'univers des données en réseau, sur le web (*Chroniques du Web*, première partie de l'ouvrage) ou ailleurs (*Etudes Cartographiques*, seconde partie). Et j'occupe encore aujourd'hui cette place indéfinie de cartographe de l'information, une place qui commence là où finissent les données et s'achève là où commence l'interprétation experte du domaine ainsi cartographié. Un espace où se sont déroulées la plupart de nos explorations de l'univers des *web data*, et dont j'ai rapporté quelques clichés regroupés dans cet ouvrage.

Chroniques du web

- Première partie -

Voir le web

«The discovery of scale-free networks induced a paradigm shift: They taught us that the many complex webs surrounding us are far from random, but are characterized by the same robust and universal architecture. I am repeatadly asked a few basic questions when I lecture about networks: Why did it take this long? Why did we have to wait until 1999 to discover the impact of hubs and power laws on the behavior of complex networks? The answer is simple: ***We lacked a map***», A.-L. Barabasi, *linked*, 2003.

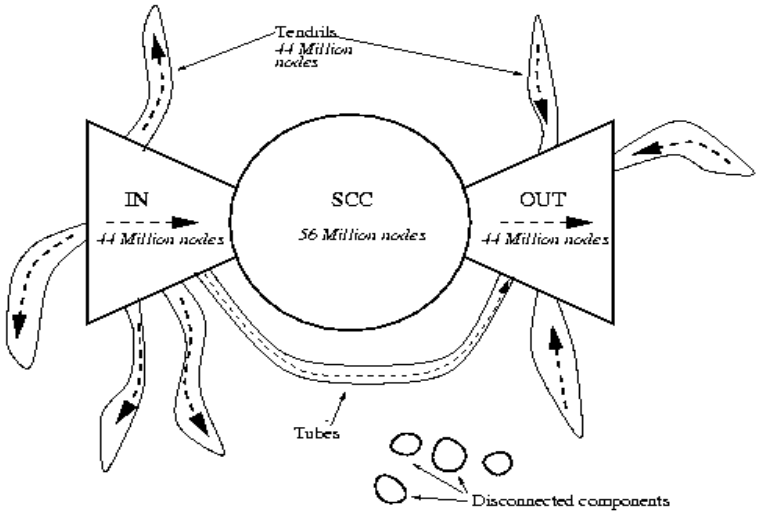


Lorsque je finis de lire en 2001 *Graph Structure In The Web*, l'article retentissant d'Andrei Broder et de ses collègues, la question s'est imposée avec évidence: le web pouvait-il avoir une *forme*? Peut-on en produire une *image* ou un *modèle* et, si oui, selon quels procédés, selon quelles méthodes, avec quels instruments? Je me souvins alors qu'il existait déjà des travaux connus sur la cartographie de l'internet comme structure physique reliant par câbles différents points du globe, ou alors comme réseau distribué de *routers* par où transitent les masses d'informations ([2]). Mais de là à imaginer que l'on puisse cartographier le web dans ce qu'il a de plus abstrait à titre d'*espace de publication de documents*, d'images et de sons, il y avait manifestement un pas qui venait d'être franchi. Une géographie documentaire...

L'article du groupe de chercheurs visait à étudier quelques-unes des propriétés statistiques du web sous la forme d'un graphe contenant plus de 200 millions d'éléments (les pages web étant considérés comme des *nodes*) reliés entre eux par plus d'un milliard et demi de liens (les liens hypertextes entre les pages ou *links*). Rien que cela! Pour ceux qui ne maîtrisent pas les mathématiques statistiques, il est difficile d'apprécier les méthodes développées dans l'article.

2 *Center for Applied Internet Data Analysis*, http://www.caida.org/home/

Mais on aperçoit rapidement que les auteurs cherchent à traduire les masses d'informations accessibles sur le web en un ensemble de propriétés physiques, comme s'il s'agissait d'un espace mesurable, gouverné par des propriétés matérielles comme le *diamètre*, la *densité*, les *composants* et leur *hiérarchie*. Autrement dit, toute une géographie d'un nouveau genre dans laquelle les liens hypertextes distribuent les documents dans des *clusters*, entre lesquels s'exercent des *forces d'attraction*, incarnés par des *flux* de liens entrants et sortants dessinant ainsi des *distances*, des *zones*, un *centre*, une *périphérie* et même des *excroissances*!



L'article est passé à la postérité pour le schéma du web «en nœud papillon» qui y figure, et l'on ne compte plus aujourd'hui les ouvrages qui le reprennent tel quel comme une sorte de «plan» général du réseau. L'idée a de quoi fasciner: plus de 200 millions de documents y sont placés dans des zones identifiables, chacune d'elles gouvernée par la façon dont les liens s'y distribuent. Autour d'un ensemble central, très dense de connexions internes, s'étirent à gauche et à droite deux «oreilles» qui comportent à peu près le même nombre de pages web, à la différence majeure que l'ensemble de gauche génère des liens qui pointent *vers* l'ensemble central et celui de droite *reçoit* des liens qui en proviennent. Le web semble massivement gouverné par une sorte de magnétisme massif, accompagné de quelques particularités remarquables comme ces excroissances (*tendrils*) sur les pôles de gauche et de droite, des *tubes* qui mènent directement de l'un à l'autre (autrement dit, on peut passer de l'un à l'autre via des chemins hypertextes de documents à documents) et quelques *disconnected components* qui posent question dès qu'on aperçoit que des «bouts du web» semblent fonctionner de façon autonome (est-ce l'effet d'un corpus parcellaire puisque ce ne sont des données partielles, un échantillon? Ou bien y-a-t-il réellement des «territoires autonomes», cachés au reste du web parce que non-connectés? Et, si oui, que recèlent-ils en termes de contenus particuliers? Peut-être des entrées sur ce que l'on appelle aujourd'hui le *dark-web*...).

Ce plan schématique avait tout d'un genre inédit de boussole, une synthèse puissante qui laissait entrevoir des propriétés essentielles de ce réseau qui engouffrait autant de données et qui commençait, en ces années 2000-2003, à tisser sa toile jusque dans nos maisons et nos objets quotidiens. Rétrospectivement, le *nœud papillon* apparaissait représenter une promesse, un projet, un défi: je compris en l'étudiant qu'il ne s'agissait pas seulement d'un exercice de synthèse où se trouvait articulée graphiquement l'essentiel des propriétés statistiques de l'étude, comme on en trouve souvent dans les publications scientifiques. Cette figure spatialisante du web annonçait

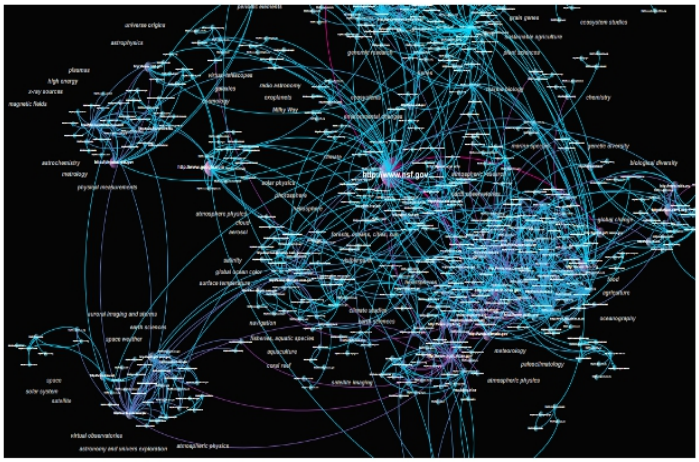
le moment où deviennent nécessaires des modèles opérationnels d'un réseau dont on commence à apercevoir les propriétés, ces «patterns robustes» ou ces «lois» qui en règle l'architecture comme l'évolution. Une fois identifiées ses propriétés essentielles, une foule de technologies pouvait être développée et mobilisée pour l'exploiter. Une source d'innovation inépuisable pour ceux qui ont commencé les premiers à dresser cette carte du réseau, les pionniers de la *Silicon Valley*. La publication de ce modèle graphique et logique du web était le premier pas de l'exploration et de l'arpentage méthodique de ces territoires numériques peuplés d'informations organisées en réseaux. Et participer à cette expédition eut tout pour moi d'une évidence, surtout que j'étais entouré de jeunes ingénieurs qui se sont plongés dans l'observation de ce vaste système d'informations distribuées qu'est le web et dont très peu de propriétés étaient connues avant les années 2000. Mais observer le web, le *voir*, n'avait rien d'une évidence: les instruments dont nous avions besoin devaient être développés et, pour partie, inventés. Le genre de *photographie* que nous espérions pouvoir produire supposait la mise en œuvre de techniques et de méthodes conçues pour un univers ouvert et distribué, et d'une croissance exponentielle. C'est ce genre de défi qu'aiment à relever parfois de jeunes ingénieurs, être en limite de la science pour ouvrir de nouvelles voies en termes de technologies dédiées aux *data*, de méthodes de calcul ou de réflexions théoriques sur la *topologie* des systèmes d'informations. De nouvelles voies, certes, mais aussi une série de décalages, voire de ruptures avec les cadres habituels à partir desquels étaient pensés jusqu'alors les systèmes d'informations ou les systèmes de connaissance.

*****

L'idée de *voir le web* fit vite office d'objectif commun de recherche technologique pour les jeunes ingénieurs qui travaillèrent au lancement de notre programme d'expérimentations. Cet enthousiasme s'expliquait par le fait qu'il devenait enfin envisageable d'explorer la structure réelle du web comme une architecture documentaire à part entière et dont nos usages quotidiens ne révèlent qu'une vue très parcellaire. Usuellement, l'accès à cet espace dépend le plus souvent de ces deux instruments que sont les navigateurs puis les moteurs de recherche qui nous renvoient leurs résultats sous forme de listes ou d'éléments ordonnés selon différents critères. Nos accès au web comme nos activités passent essentiellement par ces deux *filtres*, pratiquement transparents et naturalisés pour les internautes. Il nous fallait donc *passer de l'autre côté*, oublier le web comme *instrument* ou *vecteur* de production d'informations ou de connaissances pour en faire un *objet d'investigation*, assuré comme phénomène observable, attesté dans ses frontières matérielles. Il fallait donc réunir des instruments d'observation ou d'inspection, les imaginer et les développer si nécessaire, pour apercevoir *nous aussi* ce vaste système distribué de documents et des flux qu'y dessinent les liens hypertextes.

Certes, il s'agit encore *d'instrumentation*, donc d'un autre type de filtres tout aussi artificiels que le navigateur et le moteur de recherche. Mais cet appareillage de traitement des données serait tourné cette fois sur l'arrière cour du système, de son ossature informationnelle, des lois qui gouvernent son architecture. Et quelle architecture! Dès le départ de nos premières expérimentations, elle se révéla fascinante mais aussi déstabilisante pour des héritiers de la culture du XXe siècle et de ses instruments de prédilection comme peut l'être une bibliothèque (les livres ordonnés par dates, thèmes ou auteurs) ou une arborescence (les éléments distribués à des niveaux stricts de hiérarchie). La visualisation des connexions de document à document (ou de page web à page web) laissa rapidement apparaître un nouveau type d'architecture documentaire, une géographie foisonnante dont il allait falloir extraire des lois basées sur le nouveau principe de

proximité qu'induisait le lien hypertexte et sa distribution à grande échelle. A ce titre, le langage HTML à partir duquel est construit le document web (ou la «page») offre l'avantage remarquable d'agréger des contenus entre eux mais aussi d'insérer dans les balises des *ancres* et des *cibles* qui sont écrites, donc accessibles depuis la chaîne de caractères. On pouvait donc relever automatiquement la liste des connexions de page à page, sans aucune idée des patterns qui allaient apparaître aussitôt passés certains seuils quantitatifs. Les *web crawlers* ont ainsi joué un rôle majeur dans nos premières expéditions sur le web et ils concourent encore aujourd'hui à analyser les documents accessibles sur le réseau pour les moteurs de recherche, et à repérer la publication de nouvelles sources d'information...ou leur disparition.



Il est d'ailleurs important de souligner combien les moteurs de recherche ont contribué à traduire les données web en instruments presque «naturels» pour les usagers. Ils affichent leurs résultats sous forme de *listes* et pas de *cartes*, ils intègrent des principes de classification avec leurs catégories de recherche et restent basés sur l'extraction des «contenus» textuels, les algorithmes parmi les plus connus, comme le *pagerank*, s'inspirent des techniques d'analyse des citations en bibliométrie qui les précèdent historiquement. Ces aspects nourrissent finalement l'idée que le web est un produit de l'univers de la bibliothèque, sa *filiation* dans le monde numérique. Oui, mais le modèle d'organisation que décrivait l'article de A. Broder et ses collègues où figurait le schéma en *nœud papillon* ne correspondait à rien de connu et, si l'on pouvait tirer de ce schéma un instrument d'aide à la navigation sur le web, il n'aurait rien de commun à une liste ou à une arborescence.

Le projet d'analyser la distribution de la connectivité (parallèlement à l'indexation des contenus) constituait une voie ouverte depuis longtemps comme je l'appris petit à petit. Le *nœud papillon* ne venait en réalité que couronner une série de travaux expérimentaux entamés plusieurs années auparavant. Le principe de la mesure des espaces en différents types de réseaux à partir de la distribution des *liens* ou des *connexions* a représenté une voie grandissante du domaine du *data mining* puis du web *mining*, qui donnera naissance d'ailleurs aux *web sciences*. Dès la naissance des premiers systèmes hypertextes (avant le web), on a appliqué ainsi aux données-réseaux une série de techniques issues de la tradition de la théorie des graphes, cette branche des mathématiques appliquées qui cultivait déjà un intérêt certain pour les méthodes de projection graphique. Mais l'arrivée de données massives sur la distribution des liens hypertextes entre les pages avait permis de franchir un cap décisif: on pouvait enfin apercevoir des principes d'organisation générale ou, plutôt, des *régularités statistiques fortes* ([3]) comme certaines formes de hiérarchisation entre les documents à moyenne échelle (quelques milliers d'éléments) ou

3 B.A. Huberman, *The Laws of the Web: Patterns in the Ecology of Information*, M.I.T. Press, 2001.

encore des différences de densité hypertexte repérables dans les corpus.

Ces premières avancées illustraient déjà la nécessité d'être équipement techniquement, entouré de compétences avancées et convaincu que l'exploration du web prendrait la forme d'un lent arpentage de sa structure, comme s'il s'agissait de reconstruire le guide de la bibliothèque en parcourant soi-même, et petit à petit, tous ses rayonnages et leurs livres. L'image est à peine exagérée et résumerait presque l'essentiel des technologies d'exploration du web développées depuis les années 1990: les *crawlers* qui «sautent» de liens hypertexte en lien hypertexte pour dénombrer les pages en suivant les fils de la «toile», les bases où les contenus comme les liens sont archivés, les algorithmes de traitement de contenu et les calculs associés qui permettent de faire émerger des principes de classification, de regroupement en *clusters*, de hiérarchisation des pages selon des règles de *ranking* qui vont être retournées à l'internaute lors de ses requêtes. Nombre de ces technologies ont contribué et contribuent encore aujourd'hui à identifier les multiples propriétés du web dont certaines se révèlent surprenantes. L'efficacité vient du fait qu'elles épousent les contraintes des propriétés connues du web comme les *patterns* que forment les liens hypertextes à grande échelle: c'est par l'étude des phénomènes d'agrégation de connexions que l'on identifie des groupes ou des communautés sur le web et les réseaux sociaux. C'est en comparant les liens entre des pages web et leurs contenus (disons, les *key-words* auxquels on peut les résumer) que l'on arrive à isoler un silo d'informations pertinentes sur un sujet donné, cet *endroit* du réseau où les internautes contribuent à alimenter des connaissances collectives. Plus encore que l'information ou le document, c'est avec le lien hypertexte qu'il fallait commencer nos explorations, ce qui nous apparaissait comme la clef du voûte du système. Comme les nanomatériaux, il tisse sur différentes échelles, de la plus singulière à la plus vaste, des propriétés uniques et originales qui semblent s'emboîter pour donner naissance à des formes jusque-là insoupçonnées.

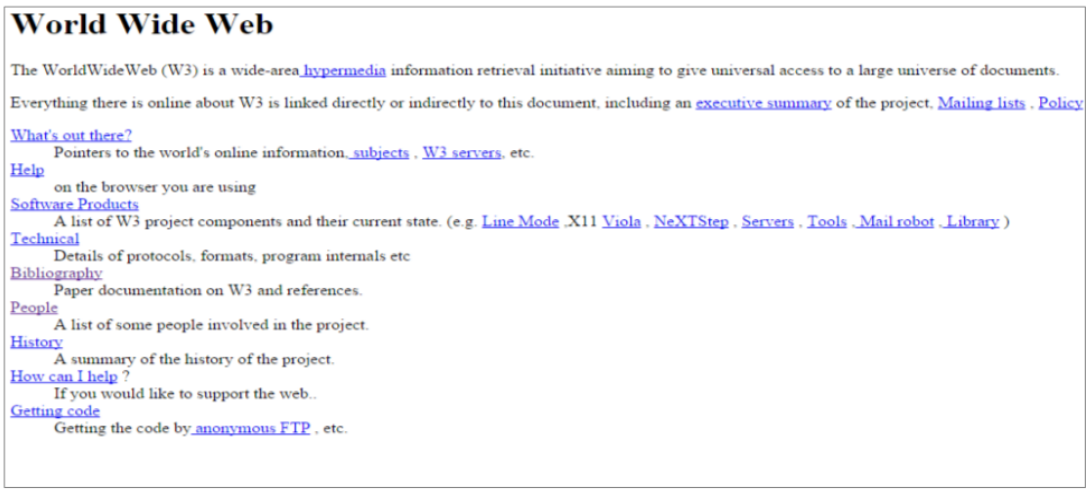
On comprend mieux alors tout l'intérêt porté au premier modèle du web en *nœud papillon* car il annonce ce que tout le monde attend: un gigantesque travail de recensement pour savoir ce qu'est donc devenu ce vaste système depuis la publication de son protocole en 1989. Ce modèle n'est pourtant pas le plan général qui permettrait de concentrer les propriétés génériques de cette architecture documentaire distribuée à grande échelle et de décrire les lois qui en gouvernent l'évolution. Pour y parvenir, il faut encore parcourir un long chemin. Mais il restera le premier modèle graphique présentant l'articulation logique de plusieurs propriétés statistiques du web comme système de documents, enfin concentrées de façon cohérente.

*****

Personne n'a encore les clefs globales du système. Chacun aimerait en détenir les plans, pas seulement par curiosité scientifique, mais aussi pour assurer l'efficience d'un service offert aux réseaux sociaux d’acteurs, pour suivre des phénomènes informationnels propagatoires comme les *rumeurs* ou les *tendances*, ou encore lancer des opérations de détection et de *tracking* de certaines thématiques ou d'entités nommées – des personnes physiques ou morales. Posséder le plan global de la distribution complète des contenus dans la structure hyperliée du système nous ferait passer de la toile indistincte à l'encyclopédie numérique universelle, contributive et raisonnée. Une façon de valider l'image développée au milieu des années 90 d'une «bibliothèque universelle et ouverte à toutes les contributions». Mais on est encore loin, à vrai dire, d'avoir isolé toutes les propriétés du web et de pouvoir le penser *comme un tout*, comme un *système*. Des pans entiers de son architecture restent encore dans l'ombre et l'objectif de pouvoir les photographier

est d'autant plus difficile qu'il faut accepter d'évoluer dans un univers qui ne semble admettre ni la clôture, ni l'exhaustivité, ni la classification hiérarchique ou arborescente. Le web n'est pas contrôlable, du moins avec les instruments classiques de gestion des systèmes de connaissances.

On peut même se demander s'il s'agit tout simplement d'un *système d'information*. La question paraît toujours aussi curieuse pour un pur produit de la technologie numérique, peuplé d'autant de machines et d'ingénieurs, un artefact culturel nourri de bases de données, d'algorithmes et d'archives, de supports réductibles à une suite de *0* et de *1*...et dont personne, en réalité, n'arrive à maîtriser réellement les masses, la diversité ou la dynamique évolutive. A bien des égards, le web reste encore un mystère. Depuis le lancement de la première page web en 1989 par une équipe du C.E.R.N., notre ignorance a grandi au fur et à mesure de son développement exponentiel, le web conservant toujours un temps d'avance sur l'observation. Tim Berners-Lee, son inventeur officiel, a même lancé le projet d'en faire une science, les *Web Sciences,* en affirmant en 2006: “Nous avons créé le Web, et nous avons pour devoir de le comprendre ". Depuis lors, nous vivons encore dans l'effervescence associée à l'arrivée d'un nouveau champ de recherche, un moment privilégié où une grande variété de disciplines viennent se confronter aux *web data*, le trésor de guerre des géants californiens de l'information. On a ainsi vu mobilisée la physique, pour étudier la logique des *flux* de liens et leurs concentrations comme des *masses*. La physique et la chimie pour essayer de comprendre les moments de mutation du système web – la *percolation* – puisqu'il ne semble pas évoluer de façon uniforme ou unilinéaire. La biologie pour le saisir comme un être vivant complexe, *auto-organisé*. On peut aussi le définir comme un *écosystème*, tout à la fois hiérarchisable sous forme d'arborescence (comme la pyramide des espèces) et décomposable en une multitude d'éléments interdépendants. On peut aussi demander à la sociologie d'expliquer les logiques sociales à l’œuvre derrière les phénomènes massifs de connexion entre des acteurs-producteurs d'informations via des documents interposés. On peut même faire appel à des réflexions plus abstraites sur le web comme *système complexe* ou encore en faire l'objet de prédilection d'une science à venir sous la forme d'une sorte d'anthropologie du numérique et des réseaux.



World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists , Policy

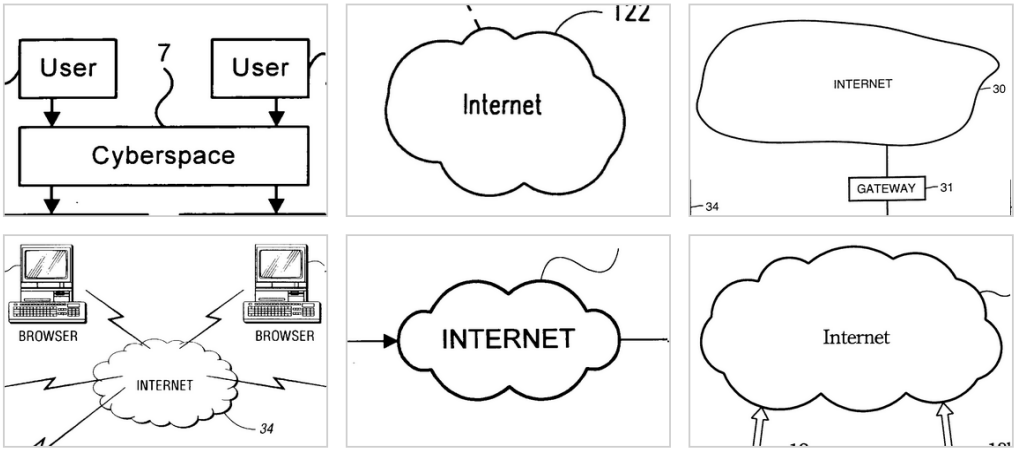
What's out there?
Pointers to the world's online information, subjects , W3 servers, etc.
Help
on the browser you are using
Software Products
A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)
Technical
Details of protocols, formats, program internals etc
Bibliography
Paper documentation on W3 and references.
People
A list of some people involved in the project.
History
A summary of the history of the project.
How can I help ?
If you would like to support the web..
Getting code
Getting the code by anonymous FTP , etc.

La première page web publiée à partir des protocoles sur lesquels a travaillé Tim Bernes-Lee depuis 1989. Le web y est défini comme un projet hypermedia ouvert et participatif: 'The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents ».

Le chemin que notre petite équipe a emprunté n'avait rien, quant à lui, d'un travail théorique et nous n'abordions pas le champ des *web data* pour valider des hypothèses préconçues. Nous avons adopté une démarche inductive en faisant le pari que les données ouvriraient sur des idées nouvelles, sans souci du genre de recherche auquel nous appartenions. Officiellement, nous ne faisions donc *rien*, les programmes scientifiques officiels et la gestion des disciplines n'autorisant

guère en France ce genre d'aventures comme celle dans laquelle nous nous sommes lancés il y a plus de dix ans. Pour certains, il s'agissait seulement d'un bricolage ingénieux de code informatique pour «faire des cartes»; pour nous, il s'agissait d'entrer de plein pied dans une branche (seulement) de ce qui allait devenir les *data sciences*. A force de lectures, de discussions et de nombreuses expérimentations, le tableau général du web a avancé par petites touches successives depuis une quinzaine d'années, en espérant y avoir contribué.

Le schéma du web en nœud papillon, comme d'autres travaux dont les nôtres, s'inscrivent dans un espace borné intellectuellement, remplissant un vide entre deux repères communs qui font les deux limites du tableau. La première c'est ce fameux nuage ([4])qui accompagne tant de documents scientifiques ou pédagogiques pour symboliser le web dont le *cloud computing* est aujourd'hui l'écho.



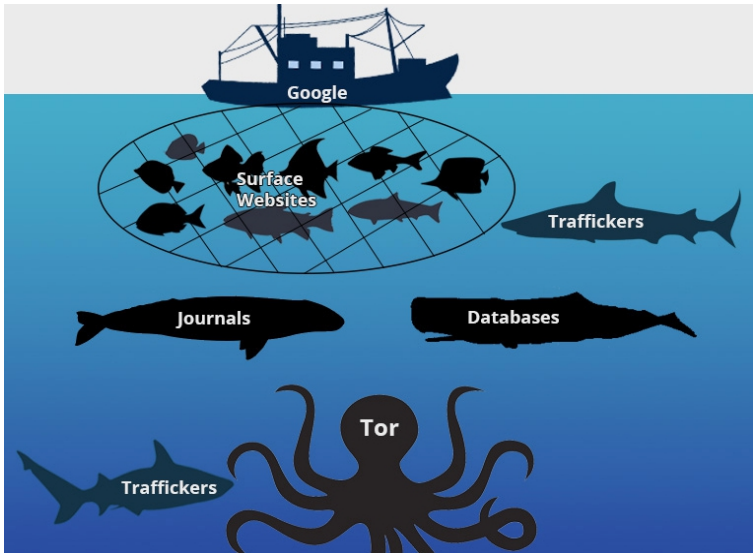
Je n'ai jamais su s'il s'agit ainsi de se dispenser des détails (on dessine le *nuage* comme on dit la «toile») ou bien d'un aveu d'ignorance, symbolisant ainsi une sorte d'extériorité incompréhensible. L'autre limite, ou borne conceptuelle est constituée par toute une lignée de travaux sur le *deep web* ([5]), au départ nommé *invisible* web ou *hidden* web ([6]). On range dans cette catégorie l'ensemble des documents ou des informations inaccessibles aux moteurs de recherche et à leur «robots» d'indexation en y mêlant pèle mêle les sites illégaux, les bases de documents formalisés (publications scientifiques, informations juridiques...), des sites gouvernementaux. Volontairement ou non, ces données à l'accès difficile semblent faiblement relié au reste du web, le web de surface tel que nous le présente les moteurs de recherche. On pourrait aussi y ranger le *dark web* autour duquel gravitent des sites de transaction commerciales illégales ou liés aux activités des *hakers*. Certains ont estimé qu'il pouvait abriter 90% de l'information du réseau mais accessibles seulement à des profondeurs insondables, et seulement avec une logistique comme du projet *Memex* ([7]) lancé en 2104 par la D.A.R.P.A. Ces profondeurs extrêmes fascinent toujours mais restent hors de portée des projets comme le notre.

4 http://noahveltman.com/internet-shape/

5 "The deep Web: Surfacing Hidden Value" appeared in The Journal of Electronic Publishing from the University of Michigan (http://www.press.umich.edu/jep/07-01/bergman.html), July 2001.

6 "Crawling the hidden web", Sriram Raghavan and Hector Garcia-molina in Proceedings of VLDB, pp.129-138, 2001.

7 http://www.darpa.mil/program/memex



C'est *là*, entre ces deux extrêmes, le nuage et le *deep-web,* que s'étend le web tel que nous le connaissons et dont il s'agit de trouver le plan, le modèle, la carte. Pour l'instant il ne s'agit que d'une esquisse mais, en traduisant patiemment les masses de données en propriétés statistiques et visuelles, le web commence à «prendre corps».