

**« L'ESPACE WEB DU SENEGAL : ÉTUDE DE SON
DEGRÉ D'OUVERTURE À TRAVERS L'ANALYSE DES
LIENS HYPERTEXTES »**

**Mémoire Master 2 de recherche Sciences de l'Information et de la Communication
Option : Nouvelles Technologies et Informations Spécialisées**

Par : GUEYE El Hadji Malick

Sous la direction de : IHADJADENE, Madjid

PRIME-CLAVERIE, Camille

DEDICACE

*A tous ceux et à toutes celles qui m'ont
manifesté leur soutien surtout dans les
moments difficiles.*

Spécialement :

A mes parents pour leur amour protecteur ;

A mes frères et sœurs, à la famille pour leur soutien indéfectible ;

Et à mes amis pour leur fidélité sans faille ;

Merci

REMERCIEMENTS

Mes remerciements vont d'abord à l'endroit de mes deux encadrants, Monsieur Ihadjadène et Madame Prime-Claverie. Sans vos conseils et suggestions mais aussi votre disponibilité et votre patience, ce travail ne verrait jamais le jour. Je remercie aussi Monsieur Perriault pour ses conseils très utiles dans l'orientation de mon sujet.

Ensuite, j'envoie un grand remerciement au Pr. Mike Thelwall du *The Statistical Cybermetrics Research Group* de l'Université de Wolverhampton pour l'aide apportée à la constitution de mon corpus de travail et au-delà pour avoir mis à la disposition des professionnels de l'information des logiciels de traitement libres et gratuits.

Enfin, je remercie tous les amis de Marseille et de Paris. C'est grâce à votre soutien moral et psychologique et vos encouragements que j'ai pu terminer ce travail.

Merci à toutes et à tous

TABLE DES MATIÈRES

Liste des figures	6
Liste des tableaux	8
Introduction.....	9
Partie I : Etat de la recherche	11
Partie II : Problématique et objectifs de recherche	15
I. PROBLEMATIQUE.....	15
I.1. <i>Enjeux des NTIC en Afrique</i>	15
I.2. <i>Vers une approche géographique de l'Internet</i>	16
I.2.1 L'émergence de la notion de cybergéographie	17
I.2.2 Le Web : entre virtualité et réalité ?	18
I.2.3 Fracture numérique, « opportunité numérique »	20
I.3. <i>L'Internet au Sénégal : état des lieux</i>	23
I.3.1 Historique.....	23
I.3.2 Les infrastructures d'accès.....	24
I.3.3 Les politiques et modalités d'accès	28
I.3.4 Evolution des sites Web sénégalais	31
II. OBJECTIFS DE RECHERCHE	32
II.1. <i>Objectifs généraux</i>	32
II.2. <i>Objectifs spécifiques</i>	32
Partie III : Revue de la littérature	33
I. METHODES QUANTITATIVES EN SCIENCES DE L'INFORMATION.....	33
I.1. <i>Définitions</i>	33
I.1.1 Bibliométrie	33
I.1.2 Scientométrie :	33
I.1.3 Infométrie.....	34
I.2. <i>Processus du traitement bibliométrique</i>	36
I.2.1 La constitution du corpus.....	37
I.2.2 Découpage du corpus en unités statistiques.....	37
I.2.3 Normalisation des données	38
I.3. <i>Analyse des citations</i>	38
I.3.1 Processus de publication : Motivations des citations.....	39
I.3.2 L'article scientifique	40

I.3.3	L'analyse du graphe de citations	41
	Facteurs d'impacts et facteurs d'influence	41
II.	DE LA BIBLIOMETRIE A LA WEBOMETRIE	42
II.1.	<i>A propos d'Internet</i>	42
II.1.1	Estimation de la taille du Web	42
II.1.2	La notion d'auto-organisation du Web	44
II.2.	<i>La webométrie</i>	45
II.2.1	Définition	46
II.2.2	Historique	47
II.3.	<i>Place des moteurs de recherche dans les études wébométriques</i>	48
II.3.1	Utilisation et limites des moteurs	48
II.3.2	Quelques réponses de professionnels de l'information	49
II.4.	<i>Analyse du graphe du Web</i>	50
II.4.1	Quelques définitions opérationnelles	51
II.4.2	Citation et « Sitation »	54
II.4.3	Le degré de connectivité du Web	58
II.4.4	La notion de Web Impact Factor (WIF)	60
	Partie IV : Analyse de l'espace Web du Sénégal	62
I.	LA CONSTITUTION DU CORPUS	62
II.	COMMENT EST STRUCTURE CET ESPACE WEB ?	67
II.1.	<i>Secteurs d'activité, types d'autorité et types de site</i>	67
II.2.	<i>Le degré d'interconnexion dans l'ensemble (.sn)</i>	75
III.	ETUDE DES HYPERLIENS EXTERNES	82
III.1.	<i>L'espace Web sénégalais et les gTLDs</i>	82
III.2.	<i>Approche géographique des liens émis par les sites sénégalais</i>	87
III.2.1	Vers la zone Afrique	90
III.2.2	Vers la zone Europe - Amérique du Nord	94
III.2.3	Vers le reste du monde	97
	Conclusion	101
	Bibliographie	103
	Annexes	110

Liste des figures

Figure 2 : Câbles sous-marins desservant l'Afrique de l'Ouest (Eric Bernard, 2002).....	26
Figure 4 : Evolution des noms de domaines .sn déclarés 1998-2002 (Source CURI)	31
Figure 5 : Evolution du nombre de sites Web. (Sources : Le Journal du Net.).....	43
Figure 7 : Terminologie de base des liens wébométriques (Björneborn, 2004)	53
Figure 8 : Connectivité du Web (Broder et al., 2000).....	58
Figure 9 : Interface de recherche de Soscibot	64
Figure 10 : Interface de restitution des résultats d'un <i>crawl</i> par Soscibot	65
Figure 11 : Répartition des différents secteurs d'activités des sites sénégalais	68
Figure 12 : Répartition des sites Web sénégalais par type d'autorité	70
Figure 13 : Réseau asymétrique entre type d'autorité et secteur d'activité	71
Figure 14 : Répartition des sites Web sénégalais par type de site.....	72
Figure 15 : Réseau asymétrique entre type de site et secteurs d'activité	73
Figure 16 : Réseau asymétrique entre type d'autorité et type de site.....	74
Figure 17 : Histogramme des plus grands sites « <i>sitants</i> » et « <i>sités</i> »	79
Figure 18 : Représentation de la connectivité de l'espace Web du Sénégal en « Bow-Tie » ..	80
Figure 19 : Réseau asymétrique entre les sites « <i>Ni sitantst, ni Sités</i> » et les types de sites ...	81
Figure 20 : Répartition des liens externes par noms de domaine génériques (gTLDs)	83
Figure 21 : Répartition des liens vers les gTLDs (.com, .org, .net, .edu, .int) par les sites sénégalais	84
Figure 22 : Graphe comparative des sites sénégalais vers les gTLDs (.com, .org, .edu, .net, .int)	85
Figure 23 : Comparaison des cinq gTLDs (.com, .org, .net, .edu, .int) par rapport aux types d'autorité et aux types de site	86
Figure 24 : Déploiement des liens émis par les sites sénégalais à travers le monde.....	88
Figure 25 : Répartition des liens émis par les sites sénégalais vers la zone Afrique	91

Figure 26 : Déploiement géographique des « <i>sitations</i> » des sites sénégalais vers les pays africains	92
Figure 27 : Répartition des 27 pays africains « <i>sités</i> » en fonction des types d'autorité	93
Figure 28 : Répartition des 27 pays africains « <i>sités</i> » en fonction des types de site.....	94
Figure 29 : Répartition des liens émis par les sites sénégalais vers l'Europe-Amérique du Nord.....	95
Figure 30 : Répartition des 30 pays européens-nord américains « <i>sités</i> » en fonction des types d'autorité	96
Figure 31 : Répartition des 30 pays européens-nord américains « <i>sités</i> » en fonction des types de site.....	97
Figure 32 : Répartition des liens émis par les sites sénégalais vers la zone Europe-Amérique du Nord.....	98
Figure 33 : Répartition des 29 pays du Reste du Monde « <i>sités</i> » en fonction des types d'autorité	99
Figure 34 : Répartition des 29 pays du Reste du Monde « <i>sités</i> » en fonction des types de site	99

Liste des tableaux

Tableau 1 : Tableau récapitulatif des grands chiffres du corpus.....	67
Tableau 2 : Répartition des sites en type de site	72
Tableau 3 : Liste des 30 plus grands sites « <i>sitants</i> ».....	76
Tableau 4 : Liste des 30 plus grands sites « <i>sités</i> ».....	77
Tableau 5 : Liste des plus grands sites « <i>sitants</i> » et « <i>sités</i> » de l'espace (.sn).....	78
Tableau 6 : Tableau récapitulatif du degré de connectivité des sites sénégalais.....	80
Tableau 7 : Répartition des liens vers les gTLDs (.com, .org, .net, .edu, .int) par les sites sénégalais	84
Tableau 8 : Répartition des cinq gTLDs (.com, .org, .edu, .net, .int) par types d'autorité et par types de sites.....	86
Tableau 9 : Répartition des liens vers les ccTLDs par zones géographiques	88
Tableau 10 : Répartition des liens vers les ccTLDs par types d'autorité.....	89
Tableau 11 : Répartition des liens vers les ccTLDs par types de site	89
Tableau 12 : Répartition des liens vers les ccTLDs par entités (« Sitants et Sités », « Seulement Sitants », « Seulement Sités », « Ni Sitants, Ni Sités »)	90

Introduction

Les nouvelles technologies de l'information et de la communication ont connu ces dernières décennies une importance toute particulière. Effet de mode, de mimétisme ou réelle révolution de la société contemporaine, force est de constater que le terme NTIC s'invite désormais dans tous les débats politiques, scientifiques, économiques, culturels (...) et intéresse particulièrement les chercheurs et les universitaires. Dans notre étude, nous voulons nous intéresser à l'aspect le plus remarquable de cette société de l'information : Internet, plus précisément le Web.

L'objectif général de cette étude est de mesurer le degré d'interconnexion des sites Web du Sénégal avec les autres sites de la toile mondiale et de déterminer ainsi leur ouverture dans ce réseau global. La motivation de ce travail est à chercher dans le retard que connaissent aujourd'hui les pays du Sud (africains particulièrement) en matière de nouvelles technologies de l'information et de la communication. Sans négliger cette fracture numérique et sans nier l'urgente nécessité de trouver des mesures pour la réduire, notre étude s'inscrit dans une démarche d'aborder autrement ce fossé notamment à travers une approche participative. Autrement dit, malgré les manques d'infrastructures et autres carences, il s'agit de surfer sur la vague de cette révolution numérique avec nos spécificités et nos richesses. Il s'agit d'exister simplement sur le Web. L'extrait suivant illustre assez bien cette vision : « ...un village branché à Internet, avec une parabole et où les femmes continuent à piler le riz à la main et à porter des seaux sur la tête sur de trop longues distances »¹.

L'existence et la participation du Sénégal sur le Web, nous voulons la découvrir à travers les relations qu'il entretient avec les autres sites du Web, et ceci en étudiant les liens hypertextes qui les unissent. Afin d'y arriver, nous allons faire appel à la wébométrie qui est une discipline héritant des techniques bibliométriques et scientométrique et qui se consacre à l'étude du contenu des pages Web, des liens hypertextes, de l'usage des sources d'information et des technologies Web.

¹ CHÉNEAU-LOQUAY, Annie. Défis liés à l'insertion des technologies de l'information et de la communication dans les économies africaines : L'exemple d'Internet au Sénégal. In : Abdelkader Djeflat et Bruno Boidin, *Ajustement et technologie en Afrique*, Publisud, avril 2002, p 103.

Ainsi, après la première partie qui sera consacrée à l'état de la recherche sur l'Internet au Sénégal suivie d'une seconde partie axée sur l'exposé de notre problématique et la définition de nos objectifs de recherche, nous allons développer dans la troisième partie la revue de la littérature sur les méthodes bibliométriques et leur cheminement vers la wébométrie et la cybermétrie. Dans la dernière partie, nous aborderons l'analyse de l'espace Web du Sénégal c'est à dire sa structure interne, son interconnexion et son « extériorisation » vers le reste du Web.

Partie I : Etat de la recherche

En général, les études qui concernent Internet abordent les technologies de connexion ou les infrastructures, la structure du réseau à travers les liens, les contenus des sites et pages Web et les usages. En Afrique, compte tenu du fossé numérique abyssal qui sépare ce continent du reste du monde, une bonne partie des études qui lui sont consacrées quant à l'insertion d'Internet est surtout axée aux enjeux des NTIC pour le développement économique et social, aux questions d'accès et d'infrastructures de télécommunication, aux usages et aux politiques gouvernementales en matière de nouvelles technologies.

Mais qu'en est-il du Sénégal ? Quelles sont les études qui ont été faites sur l'Internet dans ce pays ? Et plus précisément, en existe-t-il quelques-unes qui abordent l'analyse des liens hypertextes et le degré de connectivité des sites web sénégalais ?

Les travaux sur l'Internet au Sénégal sont relativement abondants par rapport aux autres pays de la sous région. Ceci est en grande partie dû à la « *précocité* » de son branchement aux réseaux « pré-Internet » en 1989 (le premier en Afrique de l'Ouest) grâce à l'IRD (Institut De Recherche pour le Développement, anciennement ORSTOM), de la déclaration de son nom de domaine (.sn) depuis 1992 et de sa connexion Web en 1996².

Ces études, dans leur grande majorité, s'inscrivent dans la même perspective que les thèmes énumérés plus haut concernant les pays africains.

Tout d'abord, Internet est abordé sous l'angle de ses possibles impacts dans le développement économique, d'une part, et d'autre part, de son adaptation dans les structures socio-économiques du Sénégal basées en grande partie sur l'informel (Chéneau-Loquay Annie, 2002, 2003 ; Lainé Audrey, 1999). Son insertion dans le pays est conditionnée et épouse en même temps les réalités socio-économiques et fait perdurer dans la plupart du temps les disparités géographiques entre les différentes régions du Sénégal (Guignard, Thomas, 2002) avec Dakar comme axe central.

Ensuite, les questions liées à l'accès reviennent souvent dans les études concernant le Sénégal, et l'Afrique de manière générale. Parmi ces questions, le développement des

² BRUN, Christophe. Un bref historique de l'Internet au Sénégal , IRD, juillet 2001
Disponible aussi sur l'URL : <http://www.orstom.sn/intersen/histo.shtml> [consulté le 01/03/05]

infrastructures reste le point le plus important à cause notamment du retard des pays africains dans leur globalité dans ce domaine, mais aussi du fait que ce point conditionne l'insertion et l'appropriation de l'outil Internet (Lainé Audrey, 1999, Loustau Guillaume, 2001). Eric Bernard (2003), a traité d'une manière profonde le déploiement des infrastructures Internet en Afrique de l'Ouest et a montré que le Sénégal est, parmi les pays de la sous région, le mieux, voire le plus équipé. Cet assez bon équipement, qu'il faut par ailleurs relativiser vu son retard par rapport aux normes mondiales, a permis au Sénégal d'assurer un bon maillage du territoire et de faciliter ainsi une assez bonne pénétration de l'outil Internet jusque dans les coins assez éloignés du pays (Chéneau-Loquay, Annie, 2002).

Par ailleurs, l'appropriation et le développement de l'Internet au Sénégal sont perçus aussi à travers la coopération internationale, plus particulièrement par le biais des Organisations Non Gouvernementales, ONG (Dulau Caroline, 2002). Cette « quasi spécificité » des pays en voie de développement, à cause de l'aide au développement, est fortement perceptible à Dakar qui abrite par ailleurs les sièges et les bureaux régionaux de plusieurs organismes internationaux. Ces ONG sont particulièrement actives dans l'accès à Internet aux couches de la population les plus défavorisées. Par ailleurs, elles ont été parmi les premières institutions (ex : Enda) à mettre en place leurs propres sites Web ce qui leur assure une certaine visibilité dans l'espace Web du Sénégal.

Enfin, l'espace Web du Sénégal, en tant que ensemble cohérent et évolutif, a fait l'objet de quelques études. Christophe Brun et Steven Huter (1999, 2000) chercheurs au Network Startup Resource Center (NSRC) à l'Université d'Oregon aux Etats-Unis ont essayé de dresser une topologie de l'Internet au Sénégal avec les fournisseurs d'accès et les quelques sites Web présents en cette période sur le net. La mise à jour en janvier 2000 a donné la carte suivante :

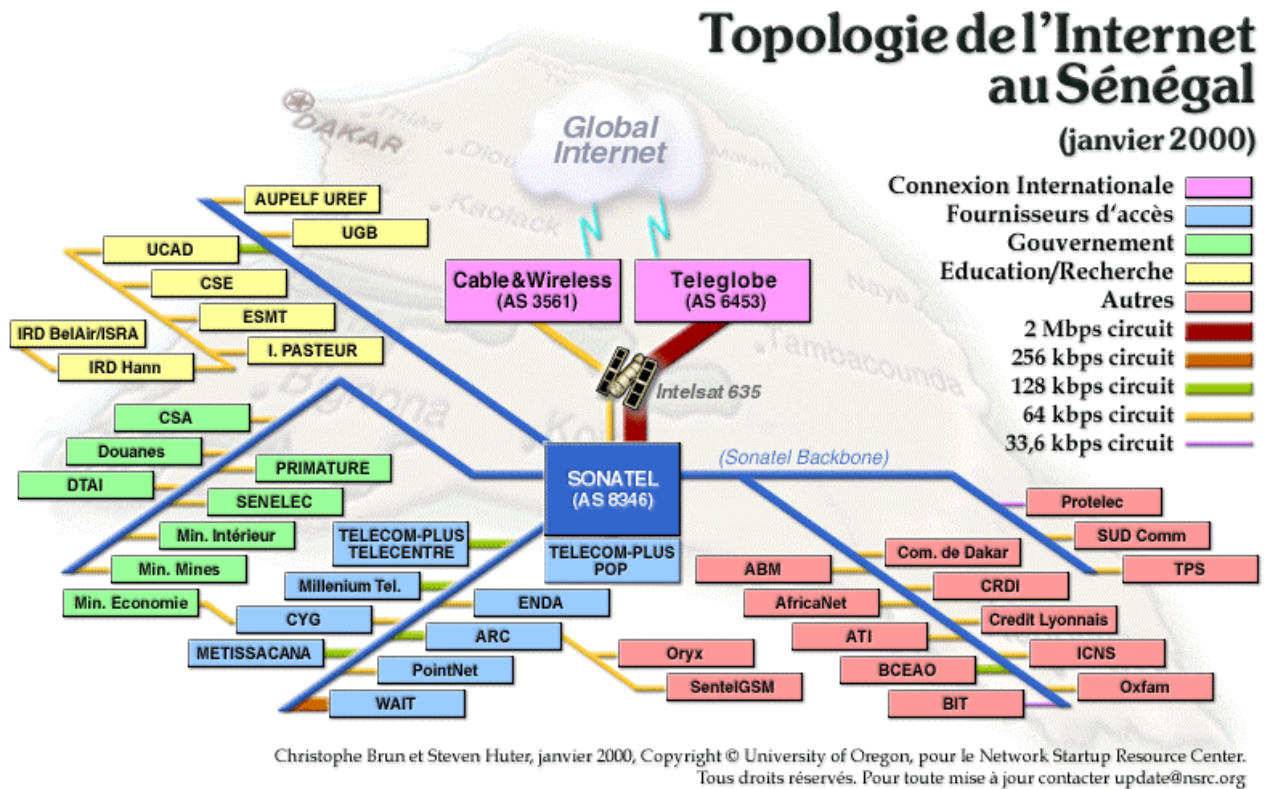


Figure 1 : Topologie de l'Internet au Sénégal (janvier 2000) (Christophe Brun, Steven Huter, NSRC)

J'ai essayé de les contacter pour avoir une carte plus récente. Ils m'ont fait savoir qu'il ne leur est plus possible d'assurer la mise à jour à cause de la prolifération des sites web sénégalais depuis cette période. Thomas Guignard (2002) quant à lui s'est penché sur le contenu des pages Web sénégalais à travers l'observation de quelques sites les plus visités comme les portails et les sites des institutions. Il a aussi tenté d'analyser les contenus des sites et les pratiques des internautes sénégalais à travers un questionnaire administré à 135 d'entre eux. Son objectif était de mesurer le degré d'extraversion du contenu des sites et des internautes sénégalais. Il a pu constater que les sites sénégalais les plus consultés présentent dans la plupart du temps des informations souvent relatives à l'Occident et que près d'un quart des internautes questionnés avouent ne consulter aucun site sénégalais !! Après quelques limites soulevées, il est arrivé à la conclusion suivante : « Une analyse des sites sénégalais mériterait d'être réalisée : malheureusement nous n'avons pas pu entreprendre une telle étude car le corpus est trop important³ ».

³ GUIGNARD, Thomas. Internet au Sénégal : une émergence paradoxale. DEA Sciences de l'information et de la communication, Université Lille 3, p.109

Comme nous venons de le voir, les études sur l'Internet au Sénégal, dans la majeure partie des cas, se sont bornées à aborder l'insertion, le développement et l'appropriation de cet outil à travers plusieurs démarches comme le développement des infrastructures de télécommunication, un accès plus élargi et plus abordable. Les enjeux et les impacts de l'Internet quant à son adaptation dans le contexte socio-économique du Sénégal, de même que le comportement des internautes sénégalais sont souvent aussi abordés. Le rôle des pouvoirs publics, des ONG et des organismes internationaux reviennent souvent dans les quelques études recensées.

A l'état actuel de notre recherche, il n'y a pas à notre connaissance (avec cependant toutes les réserves qui s'imposent) de travaux qui procéderaient à une analyse des liens hypertextes des sites Web du Sénégal et qui montreraient comment cet espace web est structuré et comment ces sites sont interconnectés entre eux et comment ils se sont liés avec le reste de la toile mondiale.

Partie II : Problématique et objectifs de recherche

I. Problématique

I.1. Enjeux des NTIC en Afrique

Si les pays africains ne parviennent pas davantage à tirer avantage de la révolution de l'information et à surfer sur la grande vague du changement technologique, ils seront submergés par elle. Dans ce cas, ils risquent d'être encore plus marginalisés et économiquement stagnants dans le futur qu'aujourd'hui". Ce passage tiré du rapport de la Banque Mondiale sur le développement d'Internet déjà en mars 1995 est sans appel. Autrement dit, il est une obligation pour l'Afrique de suivre l'évolution des NTIC, de se les approprier au risque de sombrer. Les avantages que peuvent apporter les NTIC aux pays africains sont certains. Nous n'allons pas les développer tous. On signalera juste un excellent ouvrage⁴ développé dans le cadre du programme de recherche REGARDS (unité mixte CNRS/IRD). Ce travail mené sous la coordination de Annie Chéneau-Loquay a rassemblé des chercheurs du Nord et du Sud autour des thèmes sur l'appropriation et la maîtrise des nouvelles technologies de l'information et de communication en Afrique. Aussi bien sur le plan économique, politique, social que scientifique, l'introduction des NTIC peut aider l'Afrique à venir à bout à plusieurs de ses problèmes et de quitter ainsi cette place marginale qu'elle occupe aujourd'hui au plan international. Cependant, et c'est là la particularité des travaux contenus dans cet ouvrage, il ne s'agit pas d'adhésion inconditionnelle à l'idée du « mythe de la toute puissance de la technologie ». Il s'agit plus d'analyser les voies et moyens pour tirer profit de ces outils en les adaptant aux contextes socio-économiques particuliers du continent que de considérer les NTIC comme la solution miracle qui doit permettre le développement de l'Afrique.

Notre approche de l'Internet et de son environnement global dans notre étude est aussi à recadrer dans cette perspective. Son utilité pour l'Afrique n'est plus à nier, *même si les problèmes de base, approvisionnement en eau, énergie, alimentation ne sont pas résolus*⁵. Ces problèmes de subsistance ne doivent pas empêcher une appropriation de cet outil d'information et de communication et de profiter de ses apports en terme de mise à

⁴ Enjeux des technologies de la communication en Afrique : Du téléphone à Internet. Sous la coordination d'Annie Chéneau-Loquay, Ed. Karthala, 2000. Voir :

[Hhttp://www.africanti.org/resultats/documents/enjeux.htm](http://www.africanti.org/resultats/documents/enjeux.htm)

⁵ CHÉNEAU-LOQUAY, Annie. Modes d'accès et d'utilisation d'Internet en Afrique : les grandes tendances. In : Africa e Mediterraneo, dossier Africa e il Digital Divide, n° 41, décembre 2002, p. 12-15

disposition de gisements importants d'informations qui étaient jadis inaccessibles aux pays africains. « *Les 4.000 accès du Sénégal, les 2.500 du Cameroun sont autant de fenêtres ouvertes sur les plus grandes bibliothèques scientifiques et techniques du monde, autant de points d'accès à la presse internationale, aux rapports sur les droits de l'homme, autant de vecteurs accélérant la circulation des idées* »⁶. Par ailleurs, avec Internet, l'Afrique pourrait aussi se sentir sans doute moins isolée. La visibilité mondiale qu'offre Internet peut désenclaver, culturellement et géographiquement, une bonne partie du continent. Et enfin, un autre point qui doit être une conséquence du précédent, il ne s'agira plus de se réduire au simple spectateur ou consommateur : ce qui ne fera qu'aggraver le phénomène d'extraversion constaté par Thomas Guignard⁷. Cette visibilité doit inciter à *marquer notre présence dans le monde par la production de contenus de qualité aptes à faire apprécier nos ressources et nos potentialités par l'extérieur*⁸.

I.2. Vers une approche géographique de l'Internet

La réduction des distances, le démantèlement des frontières, la relative abolition de la notion de territoire (...), voilà quelques conséquences que l'on attribue souvent à la propagation planétaire d'Internet. Cependant, la géographie, avec tout ce que cela implique comme représentation spatiale, de correspondance, de circuits d'échange des biens et services, d'interactivité entre les hommes, n'en est pas pour autant affectée, du moins dans sa signification. Seulement c'est une nouvelle géographie qui se dessine, représentée cette fois-ci par les couches physiques, les infrastructures d'accès, le trafic et les flux des données, les liens hypertextes qui interconnectent telle et telle zone... Reste à savoir quelle signification et quel sens donner à cette nouvelle géographie et ses nouveaux moyens de communication et d'échange ?

L'émergence d'une discipline qui s'intéresse à la compréhension de l'Internet comme espace à la fois virtuel et cognitif nous aidera à y voir un peu plus clair.

⁶ RENAUD, Pascal. Quand le high-tech réduit le fossé numérique. In : *Futur(e)s*, n°4, mars 2001. Disponible aussi sur l'URL : [Hhttp://www.africanti.org/resultats/documents/renaud-futur.htm](http://www.africanti.org/resultats/documents/renaud-futur.htm) [consulté le 10/03/05]

⁷ GUIGNARD, Thomas. Internet au Sénégal : une émergence paradoxale. DEA Sciences de l'information et de la communication, Université Lille 3, 180p.

⁸ SECK, Mouhamed Tidiane. Insertion d'Internet dans les milieux de la recherche scientifique en Afrique de l'Ouest. In : Enjeux des technologies de la communication en Afrique, Annie Chéneau-Loquay (dir), Karthala, 2000

1.2.1 L'émergence de la notion de cybergéographie

La cybergéographie est une nouvelle inter-discipline (à l'intersection de l'informatique, la sociologie, les sciences de l'information, la cartographie, l'urbanisme...) qui regroupe divers efforts pour étudier et représenter l'Internet et ses espaces sociaux et informationnels (Horn David, 2003)⁹. Martin Dodge, un des pionniers de cette discipline et fondateur de Cyber-Geography Research et du site cybergeography.org depuis 1997 avec ses Atlas du Cyberspaces, la définit comme : « *the geographical analysis of Internet infrastructure and usage and the spatialization and mapping of online spaces*¹⁰ ».

Basée principalement sur les techniques de cartographie et de visualisation, cette discipline s'est d'abord intéressée à l'espace physique d'Internet c'est à dire la matérialité brut du réseau comme les câbles sous-marins et les satellites. L'approche cartographique de cette partie physique du Net permet de cerner de manière pertinente le déploiement des infrastructures de télécommunications dans toutes leurs disparités et leurs discontinuités à travers le monde ; ce qui permet aussi en même temps d'évaluer la fracture numérique avec le Nord bien desservi et bien quadrillé par les câbles et les satellites, et le Sud qui présente une situation à la fois marginale et contrastée.

En plus de l'étude des flux et du « routage » des paquets de données, l'autre centre d'intérêt de la cybergéographie est d'aborder le Web en tant que espace informationnel et hypertextuel. Ceci consiste à s'appuyer sur le principe d'auto-organisation du Web pour analyser les interactions et les interconnexions et de déceler les espaces cognitifs. D'aucuns, comme David Horn¹¹, parleront de l'émergence d'une géographie « hypertextuelle ». Selon lui, *il s'agit, d'une part, de l'analyse des principes et des caractéristiques topologiques de l'interconnexion sur le Web, et d'autre part d'une tentative de « cartographier » l'information ou d'en faciliter la cognition en mobilisant des métaphores spatiales*. On retrouve dans ces tentatives de cerner l'environnement spatial du Web les mêmes principes qu'en wébométrie (voir page 58) comme la structure des liens, la théorie des graphes, l'analyse du diamètre du Web (Albert et al., 1999 ; Broder et al., 2000), la connectivité du Web...

⁹ HORN, David. [site consulté le 23/02/05].La cybergéographie : éléments pour une approche socio-spatiale de l'Internet. Essai bibliographique. Disponible sur l'URL : [Hhttp://barthes.ens.fr/atelier/geo/biblio/index.html](http://barthes.ens.fr/atelier/geo/biblio/index.html)H

¹⁰ InfoVis.net : entretien avec Martin Dodge.

Disponible sur l'URL : [Hhttp://www.infovis.net/printMag.php?num=98&lang=2](http://www.infovis.net/printMag.php?num=98&lang=2)H

¹¹ Ibid.

I.2.2 *Le Web : entre virtualité et réalité ?*

Notre approche pour cette question n'est pas de la réduire en une logique dialectique, de démontrer que le Web est soit l'un, soit l'autre. Car, vu que le caractère virtuel du Web ne fait aucun doute, nous voulons essayer de voir dans quelle mesure cette virtualité peut-elle revêtir, dans son élaboration, dans son fonctionnement ou dans ses impacts, une certaine idée de la réalité. Il s'agira d'aller chercher, au-delà des technologies de connexion, du transport des paquets de données et des liaisons hypertextuelles (à première vue instantanées et sans réels motifs), les raisons de leur élaboration et de leur donner ainsi une intelligibilité cognitive en rapport avec des considérations politiques, économiques, sociales, culturelles...

Concernant Internet dans son ensemble, (...) *en dépit de la promesse d'une ubiquité dans la connectivité, l'Internet est un réseau sélectif qui reflète la géographie physique et le développement économique*¹². Plus précisément, il apparaît clairement que le déploiement des infrastructures d'accès comme les câbles, les satellites, (bref tout ce qui compose la couche physique du réseau) est le fait d'une réelle volonté politique et obéit à des considérations économiques et financières. On est loin de la virtualité comme le soulignent Barthelemy Marc, Gondran Bernard, Guichard Eric (2003) : « *The Internet infrastructure is not virtual : its distribution is dictated by social, geographical, economical, or political constraints* »¹³. Pour illustration, les cartographies faites sur ces infrastructures au niveau planétaire donnent une vision assez nette de la fracture numérique avec les pays développés constituant le nœud de ces dispositifs et les pays du tiers monde (avec des contrastes) bénéficiant seulement de quelques ramifications. D'où les propos de Matthew Zook : « *L'Internet n'est pas en train de détruire la géographie mais connecte de manière sélective certaines personnes et certains lieux au sein de réseaux hautement interactifs, et dans le même temps en contourne largement d'autres* »¹⁴.

¹² ZOOK, Matthew. Etre connecté est une affaire de géographie. Traduit par Eric Bernard. In : *Networker*, septembre 2001, Vol 5, n°3, pp.13-17.

Disponible aussi sur l'URL : [Hhttp://www.zooknic.com/info/Zook-netWorker-2001.pdf](http://www.zooknic.com/info/Zook-netWorker-2001.pdf)H

¹³ BARTHELEMY Marc, GONDRAN Bernard, GUICHARD Eric. Spatial structure of the Internet traffic. In : *Physica A: statistical mechanics and its applications*, vol. 319, 2003, p.633-642.

Disponible aussi sur l'URL : [Hhttp://fr.arxiv.org/abs/cond-mat/0208553](http://fr.arxiv.org/abs/cond-mat/0208553)H [page consultée le 23/02/05]

¹⁴ Ibid.

S'agissant de l'univers du Web et de la virtualité proprement parlée, cette interaction (directe ou indirecte) avec la réalité est beaucoup moins évidente. Donner une quelconque intelligibilité et une signification pratique (qui s'appuieraient sur la réalité) au déploiement des sites Web, de leur interconnexion et de l'organisation hypertextuelle de la toile est un peu difficile pour la raison suivante : les raisons et motivations qui peuvent être à l'origine de la création d'un lien hypertexte sont de plusieurs sortes (voir page 54). En terme d'analogie entre bibliométrie et wébométrie, si les citations permettent, dans une certaine mesure, une représentation assez nette des relations entre centres d'intérêt, chercheurs, institutions et pays grâce notamment aux modes de fonctionnement des revues et aux règles de complication des banques de données comme Thomson ISI, on ne peut pas en dire autant pour les « *sitations* » quant à l'organisation et la compréhension du Web. Cependant, certaines études ont essayé de dépasser ces limites des liens hypertextes et de jeter un pont entre la virtualité et la réalité. Mike Thelwall¹⁵ a essayé de voir si la distance géographique entre les universités britanniques influencerait sur le degré d'interconnexion de leurs sites Web. Son étude qui concernait 109 universités est arrivée au constat suivant : plus leur distance géographique est petite, plus elles ont tendance à se « *siter* » : (...) *universities are still most likely to be linked to their neighbours*. Cependant, il a évité d'en faire une généralité à cause notamment du problème des motivations des « *sitations* » et de la relative spécificité des sites universitaires.

Par ailleurs, les gTLDs (comme .com, .org, .edu) et les ccTLDs (ex. .sn pour le Sénégal et .fr pour la France) vont permettre à ceux qui s'intéressent à la représentation spatiale du Web davantage de précision et de « fidélité » par rapport à la géographie physique. Les cartes de Martin Dodge¹⁶ montrent les différentes possibilités qu'ils offrent. La cartographie de la ville de New York par le biais de la répartition des domaines (.com.) réalisée par Matthew Zook¹⁷ en est aussi un exemple. Sa carte laisse apparaître une concentration trop importante de ces noms de domaines autour de l'île de Manhattan et Wall Street, ce qui « peut » révéler la présence d'une activité économique, financière ou commerciale assez dynamique.

¹⁵ THELWALL Mike. Evidence for the existence of géographique trends in université web site interlinking. In : *Journal of Documentation*, 58(5), 2002.

¹⁶ Voir [Hwww.cybergeography.org](http://www.cybergeography.org)H

¹⁷ [Hhttp://mappa.mundi.net/maps/maps_016/H](http://mappa.mundi.net/maps/maps_016/H) [site visité le 28/02/05]

I.2.3 Fracture numérique, « opportunité numérique »

Cette partie représente un point important à travers lequel notre étude trouve toute son essence. La fracture numérique, problématique majeure dans l'étude du déploiement et de l'utilisation des nouvelles technologies de l'information et de la communication, est aujourd'hui tellement débattue qu'elle en est presque réduite en un terme passe-partout. Aussi bien au niveau de sa définition opérationnelle, de son évaluation que des objets qu'elle tente de décrire, elle est souvent sujet à confusion. Et plus que tout autre domaine, cette nouvelle réalité de la société de l'information suit et se calque sur la géographie physique avec une nette opposition entre le Nord très en avance et le sud (particulièrement l'Afrique) très en marge de cette évolution même si des fois, il existe des configurations où *des Suds sont au Nord et des Nordes au Sud* (Annie Chéneau-Loquay, 2000).

Comment peut-on définir cette fracture numérique ?

« Que ce soit au niveau des individus, des organisations, des pays, des blocs géopolitiques, des zones géographiques, des communautés, des groupes sociaux, des métiers..., les définitions relatives à la fracture numérique renvoient à l'idée de division en deux groupes : ceux qui bénéficient de l'économie numérique (*haves*) et de l'autre, ceux qui sont exclus de l'économie numérique et de ses préposés (*have not*) »¹⁸. Donc, cette fracture désigne toujours une inégalité, une disparité dans les possibilités d'accès et les usages effectifs faits des TIC ; et ceci, quelle que soit la zone géographique, même si la disparité Nord-Sud est la plus souvent abordée notamment sous l'angle du déploiement des infrastructures d'accès.

Afin de mesurer ces disparités quant à l'accès et à l'utilisation des NTIC, des indicateurs ont été mis en place notamment par les organismes internationaux comme l'Union Internationale des Télécommunications (UIT). A part la télédensité qui décompte le nombre de lignes principales de téléphone fixe par 1000 habitants, il y a les indices dits synthétiques, plus « complets » comme *l'indice d'accès numérique* de l'UIT en 2003 qui *mesure la capacité globale des individus d'un territoire donné à accéder et à utiliser les TIC*. Cet indicateur prend en considération 5 paramètres : les infrastructures, l'accessibilité

¹⁸ RALLET Alain, ROCHELANDET Fabrice. La fracture numérique : une faille sans fondement ? In : *Dossier sur La fracture numérique. Réseaux*, vol 22, n°127-128, 2004

économique, l'Éducation, la qualité (de la bande) et l'utilisation. Il avait pour but de classer les pays en quatre catégories (excellent, bon, moyen, faible) et d'aider ainsi les pouvoirs publics dans leur politique en matière de NTIC. Cependant, la pertinence de ces différents indicateurs quant à leur capacité à quantifier et à mesurer la fracture numérique notamment dans les pays du Sud est très discutable (Annie Chéneau-Loquay, 1999 ; Richard Heeks, 2001). Ils se basent dans la plupart du temps sur des modèles et critères occidentaux comme par exemple « l'individualisme ou la personnalisation » du compte E-mail, de la ligne téléphonique, de l'ordinateur... alors que dans les pays sous-développés comme ceux d'Afrique, l'accès et les usages sont communautaires et collectifs (Pascal Renaud, 2001 ; Annie Chéneau-Loquay, 2003). « *Le critère international pour comptabiliser l'équipement téléphonique par rapport à la population, la télédensité, n'est pas un très bon indicateur en Afrique pour exprimer le service rendu...* » (Annie Chéneau-Loquay, 1999) et concernant Internet, Mike Jensen (2002) constate que, *à cause du grand nombre de comptes partagés et l'utilisation intense des services d'accès publics, il est difficile de mesurer le nombre total des utilisateurs Internet*. Pour toutes ces raisons, et sans nier le retard des pays du tiers monde, Richard Heeks (2001) ira jusqu'à affirmer que la fracture numérique est surestimée¹⁹. Il donne un exemple sur des recherches en Trinidad et Tobago où les statistiques officielles affirment qu'un foyer sur vingt est connecté au réseau alors que des études de terrains montrent qu'un foyer sur trois a accès à un messagerie électronique.

L'objectif de tous ces indicateurs est à la fois de mesurer et de tenter de réduire ce fossé numérique qui sépare notamment l'Afrique du reste du monde. Ces quelques lignes suffisent à avoir une idée sur l'état des pays africains : « *Selon les statistiques de l'Union Internationale des Télécommunications (UIT), avec 20 % de la population mondiale, l'Afrique ne compte que 2 % du réseau planétaire avec une densité globale très faible; moins de deux lignes pour 1.000 habitants en moyenne (contre 48 en Asie, 280 en Amérique, 314 en Europe - Est et Ouest - et 520 pour les pays à hauts revenus). Il est classique de dire qu'il y a autant de téléphones à Tokyo ou à Manhattan que dans toute l'Afrique sub-saharienne*²⁰ ». Les initiatives pour la réduction de cette fracture font apparaître deux courants (Rallet Alain, Rochelandet Fabrice, 2004) : l'intervention des pouvoirs publics et les lois du marché. En

¹⁹ HEEKS Richard. [site visité le 07/03/05]. La fracture numérique surestimée.

Disponible sur l'URL : [Hhttp://www.africanti.org/resultats/breves/fracturenum.htm](http://www.africanti.org/resultats/breves/fracturenum.htm)

²⁰ CHÉNEAU-LOQUAY, Annie. [site visité le 07/03/05]. Quelle insertion de l'Afrique dans les réseaux mondiaux ? Une approche géographique.

Disponible sur l'URL : [Hhttp://www.africanti.org/resultats/documents/cheneauloquay/ACL-entier.htm](http://www.africanti.org/resultats/documents/cheneauloquay/ACL-entier.htm)

Afrique, les pays du Nord et les bailleurs de fonds ont plutôt tendance à inciter à la libéralisation et à l'ouverture du marché des télécommunications. « *Face à l'énorme progression d'Internet, le risque de marginalisation des pays les moins avancés est réel. Or les pays les plus riches, plutôt que de coopérer pour installer des infrastructures, se bornent à encourager les pays en développement à s'ouvrir au marché mondial des télécommunications et à promouvoir l'initiative privée. (...) Si le démarrage de l'Internet s'est appuyé, au Nord sur une intervention massive de l'Etat, est-il sérieux de proposer aux pays les plus pauvres de faire appel au marché ?*²¹ » Résultat, la majorité des opérateurs africains se retrouve privatiser dans des conditions des fois pas vraiment les meilleures. Par ailleurs, *le cas du Sénégal est un exemple pour montrer à quel point l'idée propagée en particulier par la Banque Mondiale selon laquelle le développement d'Internet ne doit rien à l'Etat est fausse et idéologique*²². Depuis la mise en place des réseaux « pré-Internet » en 1989 jusqu'à sa mise en 1996, l'Etat sénégalais a été très présent par le biais de l'opérateur historique, La SONATEL, même si l'ouverture de son capital plus tard aux privés (France Telecom) a accéléré la diffusion des TIC au Sénégal (voir page 24).

Mais quel que soit le niveau de retard des pays africains, y a-t-il un moyen de surmonter cette fracture numérique, de participer, sans tomber dans un effet de « mimétisme », à cette société de l'information ?

L'image paradoxale *d'un village branché à Internet, avec une parabole et où les femmes continuent à piler le riz à la main et à porter des seaux sur la tête sur de trop longues distances* (Annie Chéneau-Loquay, 2002), n'est pas un « fait venu d'ailleurs » dans notre étude. Favoriser de vraies pratiques d'usage adaptées aux contextes socio-économiques et culturels locaux et transformer la fracture numérique en « *opportunité numérique* » comme souligné par le Sénégal lors du Sommet Mondial sur la Société de l'Information de Genève 2003, sont des perspectives dans lesquelles nous recadrons notre étude. Car, loin de guetter une disparition « miraculeuse » du fossé numérique du jour au lendemain et d'espérer « naïvement » des NTIC un remède à tous les problèmes de l'Afrique, et aussi dans un autre sens, de céder à un retard technologique fataliste qui peut pousser à rester au marge de la

²¹ RENAUD Pascal. Vers la désertification technologique du Sud ? In : Enjeux des technologies de la communication en Afrique, Annie Chéneau-Loquay (dir), Karthala, 2000

²² CHÉNEAU-LOQUAY, Annie. Défis liés à l'insertion des technologies de l'information et de la communication dans les économies africaines : L'exemple d'Internet au Sénégal. In : Abdolkader Djeflat et Bruno Boidin, Ajustement et technologie en Afrique, Publisud, avril 2002, p 103

révolution numérique, le Sénégal (l'Afrique) doit rester visible, s'exprimer sur le Web et saisir les opportunités éventuelles.

Mais faisons d'abord un état des lieux de l'Internet au Sénégal.

I.3. L'Internet au Sénégal : état des lieux

I.3.1 Historique

L'histoire de l'Internet au Sénégal peut se résumer en trois dates clés :

- **1989 : période « pré-internet ».** L'institut de recherche français, l'ORSTOM, qui sera renommé plus tard IRD, met en place à Dakar le RIO (Réseau Informatique de l'ORSTOM, qui changera en 1992 en *Réseau Intertropical d'Ordinateurs*), avant de l'élargir après dans la sous région. *L'objectif était d'améliorer la communication entre le siège parisien et l'ensemble de ses centres outre-mer mais aussi et surtout relier les chercheurs de l'Institut à la communauté scientifique internationale. L'échange des messages avec l'Internet global se fait via une passerelle située à Montpellier.*²³ C'était un système de messagerie de type store&forward et utilisait le protocole UUCP (Unix to Unix Copy). Notons aussi le réseau Fidonet, un autre réseau de messagerie électronique, dont le Sénégal est relié grâce à l'ONG Enda en 1992.
- **1992 : déclaration du ccTLD du Sénégal : (.sn).** Le Sénégal fait son premier pas véritable vers le réseau global Internet. Les adresses électroniques se terminant par .fr, .ca ou .org vont pouvoir être remplacées par des adresses électroniques sénégalaises, c'est à dire utilisant le ccTLD « .sn ». Ceci a été rendu possible grâce à la coopération entre l'IRD et l'Ecole Supérieure Polytechnique de Dakar. Plus tard, l'Université Cheikh Anta Diop sera chargée de gérer entièrement ce nom de domaine. Selon Eric Bernard, la déclaration de ce nom de domaine, au-delà de son importance pour l'utilisateur, peut revêtir la forme d'un véritable acte politique.
- **Mars 1996 : le Sénégal est en ligne.** Même si le premier serveur WWW d'Afrique de l'Ouest, REFER, ait été mis en ligne déjà depuis en 1995 à Dakar,

²³ BRUN, Christophe. Un bref historique de l'Internet au Sénégal , IRD, juillet 2001
Disponible aussi sur l'URL : <http://www.orstom.sn/intersen/histo.shtml> [consulté le 01/03/05]

grâce à l'Agence Universitaire de la Francophonie, le Sénégal n'entre vraiment dans Internet qu'en mars 96 lorsque la SONATEL, *l'opérateur national de télécommunication, met en place un lien Intelsat à 64 Kbps négocié avec l'opérateur MCI Worldcom et reliant le Sénégal aux USA. Le premier fournisseur d'accès grand public, Telecom-Plus, apparaît. Son premier client : la Présidence de la République*²⁴. Les anciens réseaux pré-Internet, se fondent dans un seul ensemble, l'Internet sénégalais.

1.3.2 Les infrastructures d'accès

Sur le plan des infrastructures de télécommunication, le Sénégal dispose d'un parc assez fourni et se place en position de pionnier dans la sous région et même au niveau continental.

D'abord, concernant l'accès au téléphone, le Sénégal est de très loin le pays africain qui compte le plus grand nombre de lignes publiques : 6,17 % du total des lignes contre 2,60 en Afrique du Sud, 2,90 au Swaziland²⁵. Ceci a été rendu possible grâce à une initiative originale dès 1992 : les télécentres privés. Ce sont des concessions accordées par la SONATEL (l'opérateur national de télécommunications, qui détenait le monopole sur le téléphone fixe et l'accès à l'international, monopole qui prendra fin en 2006), à des personnes privées. Ces télécentres, qu'on voit pulluler à chaque coin de rue, dans les villes comme dans les coins les plus reculés du Sénégal, sont devenus maintenant une vraie institution. Ils ont dépassé le cadre d'une simple cabine téléphonique. Ils sont des lieux de rencontre et de convivialité proposant en même temps des services de secrétariat et de dactylographie et des fois une connexion Internet, surtout à Dakar. Et selon Annie Chéneau-Loquay²⁶, cette initiative a fait que 70 % des sénégalais sont désormais accessibles par téléphone. Il faut aussi noter que le réseau téléphonique couvrant l'ensemble du territoire du Sénégal est entièrement numérique et compte plus de 2.200 km de fibre optique²⁷. Par ailleurs, la téléphonie mobile connaît une forte progression avec deux licences : Alizé, filiale à 100% de SONATEL, créée en 1996, leader du marché comptabilisait en 2001, 400.000 abonnés et 700.000 aujourd'hui ; Sentel,

²⁴ BRUN, Christophe. Ibid

²⁵ CHÉNEAU-LOQUAY, Annie. Quelle insertion de l'Afrique dans les réseaux mondiaux ? Une approche géographique. texte mis à jour : novembre 1999.

Disponible sur l'URL : [Hhttp://www.africanti.org/resultats/documents/cheneauloquay/ACL-entier.htm](http://www.africanti.org/resultats/documents/cheneauloquay/ACL-entier.htm)H
[consulté le 07/03/05]

²⁶ Ibid.

²⁷ [site visité le 22/03/05] [Hhttp://www.sonatel.sn/qui.htm](http://www.sonatel.sn/qui.htm)H

l'autre opérateur en compte près de 350.000. Un appel d'offre pour un troisième opérateur global (évoluant aussi bien sur le fixe, le mobile que sur Internet) sera lancé dans les deux mois qui viennent²⁸.

Ensuite, pour ce qui est de la connexion Internet, le Sénégal fait partie des onze pays d'Afrique où l'opérateur de télécommunications joue le jeu d'un accès universel en créant un code spécial qui permet de se connecter à Internet au coût de la communication locale dans le pays tout entier²⁹. Avec une connexion de 64Kbps dès sa mise en ligne en 1996, *le Sénégal disposait en décembre 2000 d'une bande passante à l'international de 42Mbps. Cela représente le plus gros débit à l'international d'Afrique de l'Ouest. A titre de comparaison, l'ensemble des bandes passantes des 15 autres pays de la CEDEAO (Communauté Economique des Etats de l'Afrique de l'Ouest) représente seulement un quart de ce débit*³⁰. D'aucuns, comme Eric Bernard³¹, penseront que la bande passante réelle consacrée à Internet n'était à cette période que 6 Mbps, ce qui était encore la meilleure capacité de la sous région. Le reste « *serait* » utilisé par la SONATEL pour faire passer ses appels téléphoniques. France Télécom, son partenaire stratégique depuis 1997 qui détient 42,33% du capital du Groupe SONATEL en est pour beaucoup pour cette augmentation de la bande passante notamment par son raccordement aux câbles sous-marins Atlantis 2 et SAT3/WASC/SAFE. Le câble Atlantis II relie depuis 1999 le Sénégal et le Cap Vert à l'Amérique du Sud et à l'Europe. Cette liaison de 12.000 km dessert l'Argentine, le Brésil, le Sénégal, le Cap Vert, les îles Canaries, l'Espagne et le Portugal et se connecte ensuite sur les câbles Unisur (Brésil, Argentine, Uruguay) et Columbus-2 (Italie, Espagne, Portugal, Mexique, États-Unis) déjà existants³². Le câble SAT3/WASC/SAFE (South Africa Telecommunications/West African Submarine Cable/ South Africa, Far East cable) « est le seul câble au monde à relier Nord, Sud, Est et Ouest³³ » Brian Cheesman, chargé des réseaux internationaux de Telkom, l'opérateur sud-africain. Ce câble, inauguré à Dakar le 27 mai 2003 par le Président Wade,

²⁸ Le quotidien Walfadjri, 11/04/2005. Entretien avec Monsieur Thierno Ousmane Sy, conseiller du Président chargé des nouvelles technologies. Disponible aussi sur l'URL : [Hhttp://www.walf.sn/interview/?id_inter=136](http://www.walf.sn/interview/?id_inter=136)

²⁹ CHENEAU-LOQUAY, Annie, DIOUF, Pape N'Diaye. Disponibilités et usages des technologies de la communication dans les espaces de l'échange au Sénégal. In : Enjeux des technologies de la communication en Afrique, Annie Chéneau-Loquay (dir), Karthala, 2000

³⁰ Le Sénégal décuple sa bande passante. [site visité le 07/03/05].

Disponible sur l'URL : [Hhttp://www.africanti.org/resultats/breves/SN42Mbps.htm](http://www.africanti.org/resultats/breves/SN42Mbps.htm)

³¹ BERNARD, Eric. Le déploiement des infrastructures Internet en Afrique de l'Ouest. Thèse Doctorat : Université Montpellier III, (version corr. 2004), p.218

³² BERNARD, Eric. Ibid. p.172

³³ BIDOLI, Marina. Africans now do it for themselves. Financial Mail, 07 juin 2002, [Hhttp://free.financialmail.co.za/report/telkomcable/btelkom.htm](http://free.financialmail.co.za/report/telkomcable/btelkom.htm)

est composé de deux portions : la partie africaine (SAT3/WASC) part du Portugal à Cap Town, reliant sur 14.279 km le Sénégal, la Côte d'Ivoire, le Ghana, le Bénin, le Nigeria, le Cameroun, le Gabon, l'Angola et l'Afrique du Sud. La seconde partie (SAFE), d'une longueur de 12.169 km relie l'Afrique du Sud à la Malaisie en passant par l'Inde, l'Île Maurice et la Réunion. Longtemps ignorés dans ces genres d'ouvrage, ce projet aura pour effet d'accroître de manière conséquente la connectivité internationale des pays africains et de jeter ainsi un grand pas quant à leur entrée dans les autoroutes de l'information.

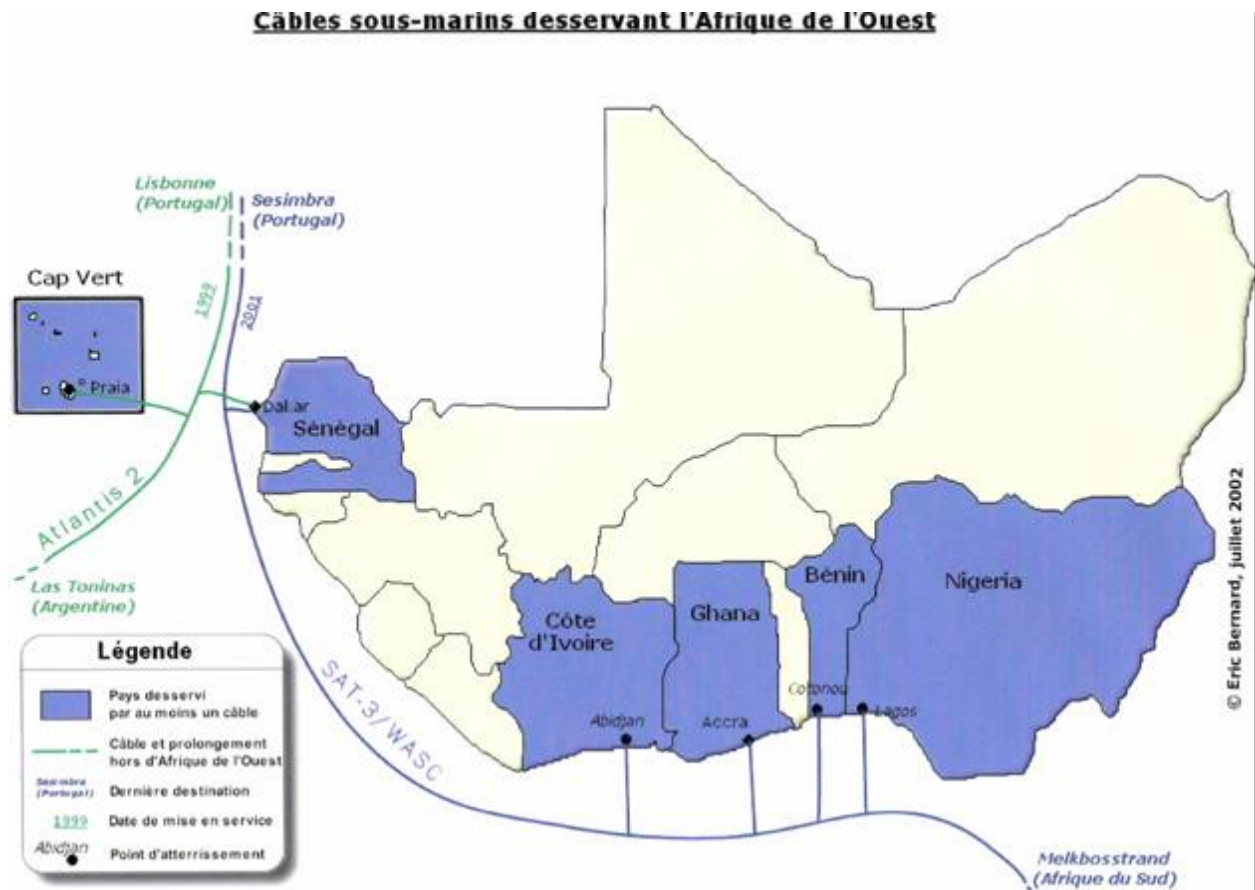


Figure 2 : Câbles sous-marins desservant l'Afrique de l'Ouest (Eric Bernard, 2002)

Ainsi, le Sénégal devrait pouvoir augmenter sa bande passante internationale avec le câble SAT3 de 42Mbps à 100Mbps³⁴. En fin 2003, elle a été de 310Mbps (155 Mbps mis en service le 14 juillet 2003 vers l'Europe sur Atlantis 2 et 155 Mbps le 30 septembre 2003 vers

³⁴ <http://www.osiris.sn/article336.html>

les USA sur SAT3/WASC/SAFE)³⁵ avant d'atteindre ½ Giga en octobre 2004³⁶. Voici l'évolution de la bande passante du Sénégal depuis sa connexion sur Internet en mars 1996.

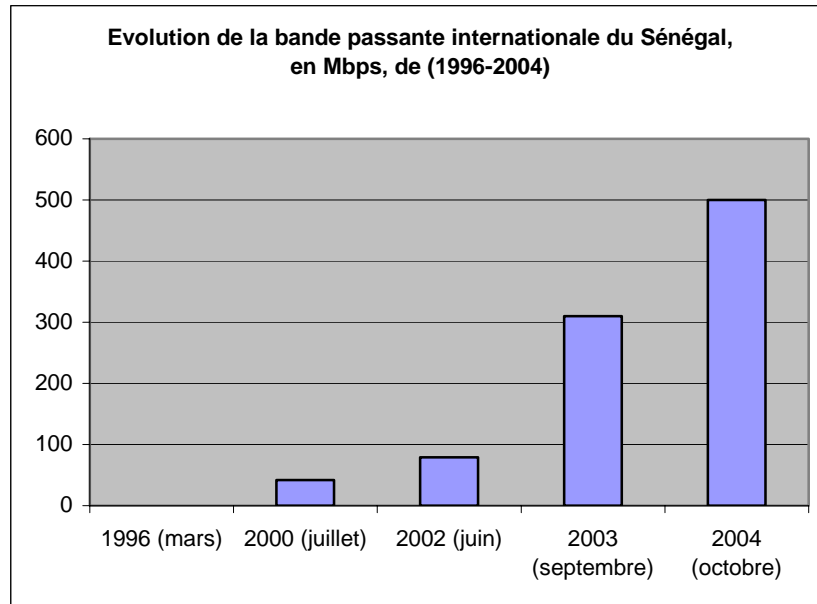


Figure 3 : Evolution de la bande passante internationale du Sénégal (1996-2004)

Cette augmentation des capacités de la SONATEL fera davantage de Dakar un « *hub* » sous régional, une plaque tournante en matière d'infrastructures de télécommunication et d'accès à Internet. Et toujours en matière d'accès Internet, la SONATEL, afin d'élargir son offre et de mieux répondre aux demandes des entreprises, des hommes d'affaires et des cybercafés (en plein essor) en matière de vitesse de navigation et de transfert des données, a lancé depuis le 03 mars 2003 la technologie ADSL devenant ainsi le quatrième pays du continent africain après l'Afrique du Sud, le Nigéria et la Tunisie à déployer cette technologie³⁷. La couverture reste néanmoins limitée à certaines zones comme la région de Dakar où la demande est assez importante. « *Avec l'ADSL, certains services de l'Internet tels que la vidéo en ligne, les catalogues virtuels en 3D, la télévision, la visioconférence via Internet, le télétravail, etc. jusque-là peu accessibles aux sénégalais, seront désormais à leur*

³⁵ SONATEL Rapport annuel 2003

³⁶ La vitesse de la bande passante Internet Sonatel portée à un demi Gigabits par seconde. OSIRIS : Revue de presse 2004. Disponible sur l'URL : [Hhttp://www.osiris.sn/article1410.html](http://www.osiris.sn/article1410.html) [site visité le 04/04/05]

³⁷ SONATEL introduit la technologie ADSL au Sénégal. Communiqué de presse SONATEL, 26 février 2003. Disponible sur l'URL : [Hhttp://www.sonatel.sn/communike/adsl.htm](http://www.sonatel.sn/communike/adsl.htm) [site visité le 21/03/05]

portée ». Et dans cette même lancée, la télévision numérique et la vidéo via la ligne téléphonique ont été testées en décembre 2004 grâce à l'appui de France Telecom et de Canal Horizons (filiale de Canal +). Six (06) chaînes sont proposées et des négociations sont en cours avec la RTS (Radiodiffusion Télévision Sénégalaise) pour inclure une chaîne nationale³⁸.

Enfin, même si toutes ces initiatives technologiques reflètent un équipement assez développé en infrastructures d'accès, la présence et la disponibilité d'un capital humain assez compétent en sont aussi pour beaucoup. *Le Sénégal se place parmi les premiers pays du Tiers monde pour le nombre d'ingénieurs et de techniciens supérieurs par rapport à sa population (...). Le pays compterait 342 ingénieurs en informatique et 467 techniciens supérieurs par million d'habitants*³⁹.

Comme remarque, nous constatons que la capacité du Sénégal en bande passante internationale dépasse largement les besoins du pays. *Cette débauche de réseaux à haut débit tournés vers l'international attire les gros clients, tel PCCI (Premium Concept Center International) qui a investi 4,5 milliards de francs CFA pour délocaliser à Dakar son centre d'appels téléphoniques, à destination de clients... européens*⁴⁰. Plusieurs autres entreprises ont investi ce secteur ; Dakar en compterait une dizaine et voudrait bien se positionner sur ce marché comme la Tunisie, le Maroc...

1.3.3 Les politiques et modalités d'accès

La question des infrastructures étant relativement réglée grâce aux efforts déployés par la SONATEL et les pouvoirs publics, reste maintenant à banaliser l'utilisation d'Internet en le rendant accessible aussi bien du point de vue de son coût que de son déploiement à toutes les couches de la population et dans toutes les régions.

³⁸ Innovation majeure en Afrique : SONATEL expérimente la Télévision numérique et la vidéo à la demande via la ligne téléphonique. Communiqué de presse Sonatel, décembre 2004.

Disponible sur l'URL : [Hhttp://www.sonatel.sn/communike/tvnum.htm](http://www.sonatel.sn/communike/tvnum.htm)H [site visité le 21/03/05]

³⁹ CHÉNEAU-LOQUAY, Annie. Défis liés à l'insertion des technologies de l'information et de la communication dans les économies africaines : L'exemple d'Internet au Sénégal. In : Abdelkader Djeflat et Bruno Boidin, Ajustement et technologie en Afrique, Publisud, avril 2002, p 103.

⁴⁰MORA, André. [site visité le 30/03/05]. Internet au Sénégal : les zones rurales sont délaissées. (janvier 2003) Disponible sur l'URL : [Hhttp://www.novethic.fr/novethic/site/dossier/index.jsp?id=31547H](http://www.novethic.fr/novethic/site/dossier/index.jsp?id=31547H)

En 2002, le Sénégal comptait 13 fournisseurs⁴¹ contre 09 en 2000⁴². Une panoplie d'offres de connexion, allant de la classique connexion commutée à l'ADSL, est aujourd'hui proposée par ces différents fournisseurs. Sonatel Multimedia, filiale Internet de l'opérateur historique, qui a lancé depuis le 15 juillet 2004 aussi des offres de connexion WIFI, représente plus de 80% de part de marché au moment où le nombre d'abonnés était estimé à 15.000 en août 2001⁴³. Autant dire que, pour les autres fournisseurs, la lutte pour la survie est rude. Il faut dire que, malgré les capacités en bande passante et les offres multiples et variées, la demande a du mal à suivre. Annie Chéneau-Loquay (2003) note un certain essoufflement de l'intérêt pour Internet à Dakar notamment. Les coûts d'accès et d'équipement en sont pour beaucoup dans ce ralentissement de la pénétration de l'Internet au Sénégal. D'une part, même si le prix de la connexion a considérablement diminué (une heure de connexion tourne aujourd'hui autour de 350 Fcfa (environ 0,5 euros) à Dakar contre 1.000 Fcfa il y a trois ans), il reste prohibitif pour bon nombre de sénégalais. D'autre part, selon Samba Sène, Directeur Général de Sonatel Multimédia " *Le principal frein au développement de l'Internet tient au prix élevé des ordinateurs. À l'exception des entreprises et d'une population de cadres, la majorité des Sénégalais n'a pas les moyens d'investir 600.000 francs CFA dans une machine neuve* "⁴⁴. Les coûts élevés incitent donc à créer des accès publics ; chose qui sera facilitée par l'existence et la bonne pénétration des télécentres dans le territoire. Comme vu plus haut, la plupart de ces télécentres offre désormais la connexion Internet à des coûts abordables. Aujourd'hui, le nombre de cybercentres est estimé à 900 dans tout le pays⁴⁵. Si l'accès et l'usage individuel dominant dans les pays développés, en Afrique, l'appropriation et l'accès aux outils de communication sont essentiellement collectifs étant donné le faible niveau de vie moyen des populations comparé au coût du matériel et de la communication elle-même (Chéneau-Loquay, Annie, 2003). Et pour Pascal Renaud⁴⁶ « *L'accès collectif est sûrement la solution la mieux adaptée lorsqu'il s'agit de répartir des moyens limités. Et Internet s'y prête : les PC regroupés en grappe dans des cybercentres partagent les frais de connectivité* ».

⁴¹ Internet au Sénégal : Liste des fournisseurs d'accès Internet.

Disponible sur l'URL : [Hhttp://www.orstom.sn/intersen/isp.html](http://www.orstom.sn/intersen/isp.html)H [site visité le 29/03/05]

⁴² BRUN, Christophe, HUTER Steven. Topologie de l'Internet au Sénégal. Université of Oregon, NSRC, janvier 2000

⁴³ JENSEN, Mike. [site visité le 28/02/05]. African Internet Connectivity.

Disponible sur l'URL : <http://www3.sn.apc.org/africa/afrmain.htm>

⁴⁴ Internet au Sénégal : les zones rurales sont délaissées. (janvier 2003)

Disponible sur l'URL : [Hhttp://www.novethic.fr/novethic/site/dossier/index.jsp?id=31547](http://www.novethic.fr/novethic/site/dossier/index.jsp?id=31547)H [site visité le 30/03/05]

⁴⁵ Quotidien Walfadjri, 24 mars 2005.

⁴⁶ Ibid.

Dan cette même perspective, beaucoup d'initiatives vont être développées pour permettre l'appropriation de l'Internet par les populations les plus défavorisées et les plus enclavées. Les Centres Multimédias Communautaires (CMC), développés par les pouvoirs publics avec l'appui de l'UNESCO en sont un exemple. Ce projet part du constat sur la disparité entre centres urbains et campagnes en matière d'accès aux nouvelles technologies de l'information. Car, il existerait une vraie fracture numérique entre régions. Par exemple, sur les 184 cybercentres recensés par Thomas Guignard⁴⁷ en 2001 dans son étude, 111 se trouvent dans la région de Dakar, concentrant ainsi 60 % des cybercentres sur 0,3% du territoire avec 25 % de la population totale du Sénégal. D'où le constat suivant : *Internet est d'abord l'apanage des centres villes et de leurs élites mieux reliées aux centres mondiaux qu'à leur propre hinterland...*⁴⁸ Ce projet va donc donner la priorité aux zones rurales et périurbaines. L'objectif des CMC est ainsi de favoriser l'appropriation des NTIC aux citoyens les plus défavorisés et de *faire progresser le niveau de connaissance des populations sur les problèmes de leur terroir, de leur pays et de l'étranger*⁴⁹. Une autre initiative et non des moindres est la signature, le 25 octobre 2004 à Dakar, d'un protocole d'accord entre le ministre de l'Education, Moustapha Sourang, et le PDG de Microsoft Europe, Moyen-Orient et Afrique, Jean-Philippe Courtois, portant sur l'accès à Internet de trois millions d'élèves et étudiants sénégalais⁵⁰. Selon les propres termes du Ministre : *"Grâce à cet accord, trois millions d'élèves et étudiants vont bénéficier de l'accès à une machine et à Internet et 60% des bacheliers pourront exercer un métier lié à l'informatique"*, à travers notamment l'acquisition de 10.000 ordinateurs et la formation de plus de 2.000 professeurs. Et enfin, la célébration de la fête de l'Internet est aussi l'occasion pour les organismes impliqués dans le développement des NTIC d'élargir l'« @lphabétisation » des populations. Pour l'édition 2005, qui se déroulait du 20 au 27 mars 2005, le Forum des Cybercentres du Sénégal (FOCYS), a organisé des journées portes ouvertes en offrant gratuitement 30 minutes de connexion à tout le monde⁵¹. Ceci dans le but de permettre aux internautes, surtout aux néophytes, de découvrir les services comme la messagerie électronique, les forums de

⁴⁷ Ibid.

⁴⁸ CHENEAU-LOQUAY, Annie. Modes d'accès et d'utilisation d'Internet en Afrique : les grandes tendances. In : *Africa e Mediterraneo*, dossier *Africa e il Digital Divide*, n° 41, décembre 2002, p. 12-15

⁴⁹ Centres Multimédias Communautaires : Ouvrir le monde rural à l'Internet. Revue de presse OSIRIS, 2004. Disponible sur l'URL : [Hhttp://www.osiris.sn/article1349.html](http://www.osiris.sn/article1349.html)H [site visité le 04/04/05]

⁵⁰ Trois millions d'élèves et étudiants bientôt à l'école de l'Internet. OSIRIS, revue de presse 2004.

Disponible sur l'URL : [Hhttp://www.osiris.sn/article1373.html](http://www.osiris.sn/article1373.html)H [site visité le 05/04/05]

⁵¹ Fête de l'Internet : 30 minutes de connexion gratuite, mais un bilan mitigé. OSIRIS, revue de presse 2005. Disponible sur l'URL : [Hhttp://www.osiris.sn/article1673.html](http://www.osiris.sn/article1673.html)H [site visité le 05/04/05]

discussion et l'initiation à la recherche. "L'étape la plus difficile, c'est la première entrée dans un cyber-café. Après, c'est une drogue..." dira tout simplement Amadou Moutar Sow président de FOCYS.

1.3.4 Evolution des sites Web sénégalais

Vu la rareté des études effectuées sur ces sites Web, très peu d'informations sont aujourd'hui disponibles à leur sujet. L'étude que nous sommes en train de mener, nous l'espérons, approfondira davantage la connaissance de cet espace Web et permettra de mieux le comprendre aussi bien du point de vue de sa structure que de son degré d'ancrage dans le réseau mondial.

Le nombre des noms de domaines (.sn) enregistrés, comme partout ailleurs dans le monde, a connu une évolution rapide. D'après les statistiques de la Commission Université Réseaux d'Informations (CURI), organisme rattaché à l'Université Cheikh Anta Diop de Dakar et chargé de l'enregistrement et de la gestion des noms de domaines (.sn), les sites sénégalais déclarés sont passés de 62 en 1998 à 914 en 2002.

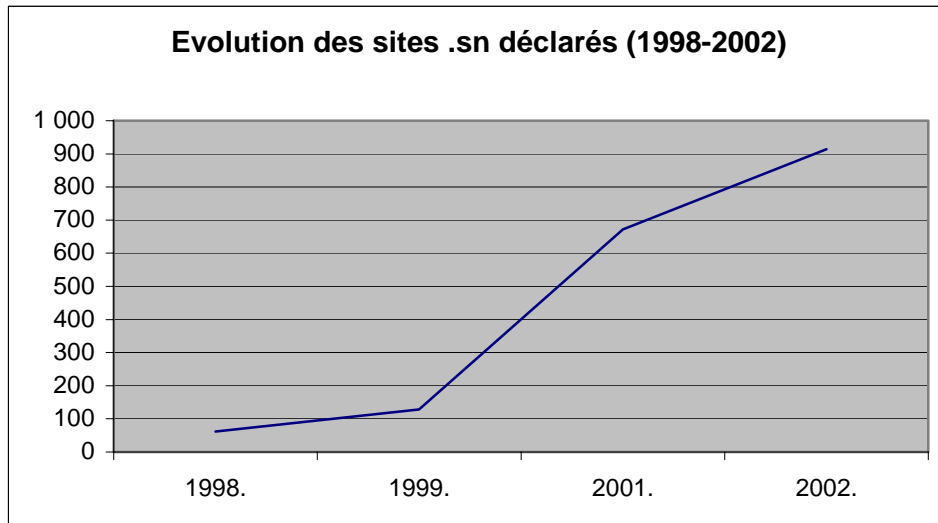


Figure 4 : Evolution des noms de domaines .sn déclarés 1998-2002 (Source CURI)

Mais selon Thomas Guignard⁵², il existerait une grande différence entre les sites déclarés et ceux étant effectivement en ligne. Par exemple, en 2001, alors qu'on dénombrerait

⁵² Ibid.

672 sites déclarés, ils n'étaient que 160 à être en ligne. Les prix assez prohibitifs de la création d'un site Web expliquent peut-être ce problème.

II. Objectifs de recherche

II.1. Objectifs généraux

L'objectif principal de cette étude est de mesurer le degré d'interconnexion des sites Web du Sénégal, les noms de domaines (.sn) plus précisément, avec les autres sites de la toile mondiale et de déterminer ainsi leur visibilité dans ce réseau global. Ce travail commencera par la constitution d'un corpus regroupant l'ensemble des noms de domaines (.sn) en ligne. Et à travers les méthodes wébométriques comme l'analyse des liens, nous comptons arriver à déceler les liens externes à l'espace Web du Sénégal, c'est à dire les liens partant de cet ensemble vers des sites « non sénégalais ». Ce qui nous permettra d'analyser comment le Sénégal « s'externalise » sur le Web et à quel degré.

Afin de bien l'atteindre, cet objectif principal est assorti d'objectifs secondaires ou spécifiques qui nous permettront de bien le préciser dans son élaboration et sa réalisation.

II.2. Objectifs spécifiques

- **Mesurer la taille de l'espace Web du Sénégal :** ce sera le point de départ de cette étude. Comme nous l'avons vu plus haut, les noms de domaines (.sn) déclarés et enregistrés auprès de la CURI NIC Sénégal diffère largement de celui des sites effectivement en ligne.
- **Structurer cet espace Web :** nous comptons aussi catégoriser les sites sénégalais en domaine d'activité, en type d'autorité et en type de site pour ensuite étudier leur interconnexion.
- **Lister les liens internes et externes :** pour précision, ce sera les liens internes à l'ensemble Web du Sénégal, les liens qui sortent de cet ensemble vers d'autres sites de la toile mondiale.
- **Identifier et stratifier les zones géographiques** vers lesquelles pointent les sites sénégalais grâce à l'identification des ccTLDs.

Partie III : Revue de la littérature

I. Méthodes quantitatives en sciences de l'information

I.1. Définitions

"Pourquoi ne pas appliquer à la science ses propres instruments ? Pourquoi ne pas mesurer, généraliser, faire des hypothèses, tirer des conclusions" se demandait Derek John de Solla Price dans son célèbre livre *Little Science, Big Science* (1963).

Cette citation nous permet d'entrer dans la partie préliminaire de notre étude et qui est consacrée aux méthodes quantitatives : scientométrie, bibliométrie et infométrie. Les travaux de De Solla Price ont été particulièrement déterminants notamment en scientométrie : "*The key figure in this new quantitative studies was Price, whose writings, especially **Little Science, Big Science** had a major impact on thinking about the growth and evolution of scientific journals*"⁵³. Ces outils permettent en somme : d'évaluer le travail d'un chercheur, de mesurer l'évolution d'un domaine de recherche, d'évaluer l'impact d'un article et le prestige et la qualité d'une revue...

Dans cette partie, nous tenterons de rapporter les différentes définitions qui ont été données à ces méthodes, leurs spécificités et les contextes qui ont prévalu à leur développement.

I.1.1 Bibliométrie

La bibliométrie est définie en 1969 par Pritchard comme *l'ensemble des méthodes et techniques quantitatives - de type mathématique/statistique - susceptibles d'aider à la gestion des bibliothèques et d'une manière très générale des divers organismes ayant à traiter de l'information*.⁵⁴

I.1.2 Scientométrie :

Pour Xavier Polanco (1995), on peut considérer la scientométrie comme la bibliométrie spécialisée au domaine de l'IST (l'information scientifique et technique). Toutefois, la

⁵³ MEADOWS, A. J. "Theory in information science", *Journal of Information Science*, vol. 16, n° 1, 1990, p. 59-63

⁵⁴ NOYER, Jean-Max. [site consulté le 15/01/05]. Scientométrie, infométrie : pourquoi nous intéressent-elles ?. Disponible sur l'URL : [Hhttp://biblio-fr.info.unicaen.fr/bnum/jelec/Solaris/d02/2noyer_1.html](http://biblio-fr.info.unicaen.fr/bnum/jelec/Solaris/d02/2noyer_1.html)

scientométrie désigne d'une manière générale l'application de méthodes statistiques à des données quantitatives (économiques, humaines, bibliographiques) caractéristiques de l'état de la science.

Une petite comparaison entre ces deux termes permet de détecter que, bien qu'ils se basent tous sur les mêmes techniques et méthodes (voir page 35) à quelques différences près, ils ont des objets d'étude différents et visent de ce fait des objectifs différents. Ces propos de Brookes résument tout : « Alors que la bibliométrie aurait pour objet d'étudier les livres et les revues et pour objectif de comprendre les activités de la communication de l'information, la scientométrie aurait pour objet l'étude des aspects quantitatifs de la création, la diffusion et l'utilisation de l'information scientifique et technique et pour objectif la compréhension des mécanismes de la recherche comme activité sociale »⁵⁵.

Ces propos peuvent être représentés sommairement comme suit :

Bibliométrie -----> bibliothéconomie -----> étude descriptive

Scientométrie -----> science de la science -----> étude sociologique

I.1.3 Infométrie

Plus récent, ce terme a été adopté en 1987 par la F.I.D. (Fédération Internationale de Documentation). Tague-Sutcliffe (1992) le définit comme : *"the study of the quantitative aspects of information in any form, not just records or bibliographies, and in any social group, not just scientists"*. L'infométrie devient l'ensemble des activités métriques relatives à l'information, couvrant ainsi aussi bien la bibliométrie que la scientométrie⁵⁶. On retrouve cette même conception chez Le Coadic⁵⁷, pour qui, l'infométrie regroupe, en plus de la bibliométrie et de la scientométrie, la médiamétrie, la muséométrie et la wébométrie. Ceci dit, l'amalgame pour désigner ces trois termes est fréquent (Lafouge, Boukacem, 2004).

Polanco (1995) résume assez bien ces trois concepts : *"Les études quantitatives de la science et de la technologie représentent le champ de recherche où l'on utilise les méthodes et*

⁵⁵ BROOKES, B. C. Biblio-, sciento-, info-metrics ??? What are we talking about? First International Conference on bibliometrics and theoretical aspects of information retrieval, August 24-28, Belgium, 1987

⁵⁶ PLANCO, Xavier. [site consulté le 15/01/05]. Aux sources de la scientométrie.

Disponible sur l'URL : [Hhttp://biblio-fr.info.unicaen.fr/bnum/jelec/Solaris/d02/2polanco1.html](http://biblio-fr.info.unicaen.fr/bnum/jelec/Solaris/d02/2polanco1.html)

⁵⁷ LE COADIC, Yves François. [site consulté le 13/01/05]. Infométrie mathématique et infométrie statistique.

Disponible sur l'URL :

[Hhttp://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/03/63/sic_00000363_03/sic_00000363.html](http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/03/63/sic_00000363_03/sic_00000363.html)

les techniques mathématiques, statistiques et de l'analyse des données en vue de rassembler, manipuler, interpréter et prévoir une variété de caractéristiques telles que la performance, le développement et la dynamique de la science et de la technologie"

A titre indicatif, nous utiliserons, tout au long de ce travail, plus souvent le terme bibliométrie pour désigner l'ensemble des activités de métriques, et ceci par pur souci de commodité.

Mais quels que soient le domaine ou la discipline auxquels on peut les rattacher, quels que soient leurs objets d'étude et leurs objectifs, ils se fondent tous sur les mêmes lois de distributions statistiques. Meadows (1990) nous signale que l'intérêt pour les caractéristiques quantitatives de l'information, c'est-à-dire pour une approche de type bibliométrique, s'est particulièrement développé à partir des années 1950, sous l'impact du travail de Shannon (1949), ayant comme fondement les lois bibliométriques à savoir la loi de Lotka (1926) concernant les auteurs, la loi de Bradford (1935) concernant les sources d'information et la loi de Zipf (1936) qui concerne les mots d'un lexique ou d'un discours.

- La loi de Lotka a pour objectif de mesurer la contribution de chaque chercheur au progrès scientifique.
- La loi de Bradford a pour visé la gestion des abonnements et de manière plus précise de connaître le « noyau » des périodiques d'un domaine.
- La loi de Zipf vise l'étude linguistique des écrits littéraires par le biais de la fréquence d'utilisation des mots.

Ces trois lois, comme pour toutes lois hyperboliques, sont caractérisées par un faible cœur et une forte dispersion (Prime-Claverie, 2004). Le cœur représente un petit nombre d'éléments ou d'individus ayant une forte fréquence. En d'autres termes, cela veut dire que peu de revues publient la grande majorité des articles (loi de Bradford), peu de mots sont très fréquents dans les textes (loi de Zipf), peu d'auteurs publient beaucoup (loi de Lotka). La dispersion caractérise un très grand nombre d'éléments ou d'individus ayant une faible fréquence. Ce qui revient à dire aussi que la grande majorité des revues ne publient qu'une infime partie des articles (loi de Bradford), la plupart des termes n'apparaît qu'une seule fois dans les textes (loi de Zipf), la majeure partie des chercheurs ne publie qu'un seul article pour une période donnée (loi de Lotka).

Ces lois ont servi de fondement au développement, plus tard dans les années 60, des méthodes quantitatives comme la scientométrie et dont Price (analyse des citations) sera la figure emblématique. Mais ces dites méthodes citationnistes n'ont été largement utilisées qu'avec l'arrivée des outils développés par l'Institute for Scientific Information (maintenant Thomson ISI) et des recherches de son fondateur, Eugene Garfield⁵⁸. Meadows (1990) nous apprend que : “*One important area of Price's work covered the applications of citation analysis. In this, he relied on the contemporaneous activities of Garfield in developing the concept of a citation index*”

Les travaux de Garfield ont donné naissance à des outils devenus quasi incontournables dans les traitements bibliométriques, notamment en analyse des citations, aussi bien pour la compréhension de la production, la diffusion des écrits et la composition de la communauté scientifique et les liens qu'entretiennent ses membres. Ces outils sont les banques de données Science Citation Index (SCI), Social Science Citation Index (SSCI) et le Arts and Humanities Citation Index (AHCI), mais aussi le Journal of Citation Report (JCR) qui donne le facteur d'impact des revues .

Nous reparlerons de ces banques de données et de la notion de citation tout au long de la prochaine partie qui est consacrée aux différents stades que doit suivre un traitement bibliométrique.

I.2. Processus du traitement bibliométrique

Les études bibliométriques travaillent sur des corpus volumineux de publications scientifiques, généralement des articles primaires ou des brevets et suivent en général plusieurs étapes passant de la constitution du corpus jusqu'à l'interprétation des résultats de l'analyse. Vu l'orientation que nous comptons donner à notre étude et le caractère spécifique de ces genres d'étude (qui essaient d'appliquer les techniques bibliométriques à l'environnement web), nous abordons dans cette partie les trois étapes les plus problématiques dans une étude wébométrique pour finir avec une notion très capitale aussi pour notre recherche c'est à dire *l'analyse citationniste*. Les trois étapes que nous aborderons sont :

- La constitution du corpus
-

⁵⁸ ARCHAMBAULT, Eric, VIGNOLA G., Étienne. L'utilisation de la bibliométrie dans les sciences sociales et les humanités. Conseil de recherche en sciences humaines du Canada (CRSH). Rapport final, août 2004.

- Découpage du corpus en unités statistiques
- Codification des unités statistiques

1.2.1 La constitution du corpus

La constitution du corpus commence par la collecte des données sur lesquelles va porter l'étude donnée. Ce sont les banques de données bibliographiques de l'ISI, entre autres banques de données, qui sont le plus souvent utilisées pour constituer ces corpus. Ceci est dû au fait qu'elles présentent beaucoup d'avantages par rapport aux autres banques de données (Katz, Hicks, 1998) :

- Elles présentent une très bonne couverture des domaines de recherche dans la mesure où elles recensent systématiquement, avec cependant quelques biais, tous les articles et les thèmes des revues qu'elles couvrent.
- Le critère d'inclusion d'une revue dans le SCI, SSCI et le AHCI est le nombre de citations qu'elle reçoit, ce qui rejoint les travaux de De Solla Price (1963) « *le degré d'utilisation semble être un meilleur test de qualité* » ; au lieu d'une approche basée sur la quantité des articles publiés,
- Elles contiennent les adresses institutionnelles des auteurs d'un article spécifique, très important pour l'analyse de la collaboration.
- Seules les banques de données de Thomson ISI contiennent les citations. Ces informations permettent de mesurer l'impact de la recherche. Katz et Hicks (1998) considèrent que cette caractéristique justifie à elle seule l'usage de ces banques de données comme outil de politique scientifique et de gestion de la recherche.

Les banques de données de Thomson ISI possèdent aussi certains désavantages qui tiennent au fait qu'elles sont relativement coûteuses et ne se prêtent pas aussi bien en recherche en sciences sociales qu'en sciences naturelles (Archambault et Vignola, 2004).

1.2.2 Découpage du corpus en unités statistiques

Cette étape est aujourd'hui moins fastidieuse avec les efforts considérables que fournissent les serveurs de banques de données dans la compilation des références. Les notices bibliographiques sont des ensembles structurés d'information composés de champs comme : auteurs, *titre*, *mots-clés*, *date de publication*, *langue*, *résumé* ... Chaque champ est composé d'un nom de champ et d'un contenu. « Certains champs sont particulièrement riches

d'information pour contribuer à l'analyse de l'univers scientifique. Les champs *mots-clés* et *titre* en sont de bons exemples. Ils figurent d'ailleurs parmi les champs les plus souvent utilisés dans les études bibliométriques » (Prime-Claverie, 2004).

1.2.3 Normalisation des données

La normalisation du corpus est une étape très importante, car elle conditionne pour une grande partie la bonne analyse des données collectées. Malgré les efforts déployés par les serveurs pour l'harmonisation des références, certains champs posent beaucoup de problèmes dans le cadre d'un traitement bibliométrique comme le champ *adresse des auteurs* (Archambault, Vignola., 2004), qui présente souvent beaucoup de variances. Toujours selon eux, il faut noter que les banques de données sont optimisées pour retracer des articles plutôt que pour faire des calculs complexes de dénombrement. En d'autres termes, elles sont conçues pour des usages bibliographiques plutôt que bibliométriques. Le travail de bibliométrie commence donc avec le conditionnement de données bibliographiques dans le but de constituer des banques de données bibliométriques. Le travail consiste principalement à normaliser les données. Donc tout ceci nécessite un travail de nettoyage, d'épuration et d'harmonisation du corpus (ajout ou suppression de champs) pour arriver à un bon niveau de traitement.

Ces différentes étapes ainsi présentées, même si elles posent de temps en temps des problèmes dans le cadre d'une étude bibliométrique, elles sont largement facilitées par les efforts des serveurs de banques de données en matière de compilation et d'harmonisation des références bibliographiques. Dans notre contexte d'étude, vu la spécificité et l'hétérogénéité des documents web, ces étapes, surtout celles concernant le découpage et la codification du corpus, sont assez fastidieuses comme nous le verrons plus loin dans la troisième partie.

I.3. Analyse des citations

L'analyse des citations, malgré quelques limites, va fortement bouleverser les méthodologies d'analyse des écrits scientifiques de même que la compréhension de la sociologie des sciences.

1.3.1 Processus de publication : Motivations des citations

Pour comprendre les motivations qui peuvent pousser un chercheur à citer ses pairs dans ses travaux, il faut garder en tête que la connaissance scientifique objective est cumulative par essence. Chaque nouvelle connaissance scientifique enrichit, modifie, perfectionne ou réfute totalement dans certains cas, la connaissance précédente. Cette caractéristique de cumul est partagée par la littérature scientifique. Dans la pratique, la citation n'est rien d'autre que la relation qui lie un document citant et le document cité. Price (1970) précisera davantage cette notion de citation : « Si l'article A a une note bibliographique utilisant et décrivant l'article B, alors A contient une référence à B, et B reçoit une citation de A ».

Et pour histoire, il est d'usage depuis le XIX^{ème} siècle que le chercheur mentionne à la suite de son article l'ensemble des travaux qui l'ont aidé dans le cadre de sa recherche. Ces citations permettent d'une part, aux lecteurs de consulter les travaux qui ont inspiré l'auteur ; d'autre part, c'est aussi une façon pour lui de rendre hommage à ses prédécesseurs. Selon Case et Higgins⁵⁹, il existerait deux écoles pour étudier les motivations des citations : la première considère la citation comme une dette intellectuelle vis-à-vis des pairs qui ont inspiré le chercheur. Et l'autre pense que la citation sert avant tout les intérêts de l'auteur puisqu'il cite pour rendre son article beaucoup plus crédible, beaucoup plus persuasif.

Ainsi, vu que le monde scientifique forme une communauté qui ne cesse de s'élargir et où chaque nouveau savoir vient se raccorder à ceux existant, on est à même de comprendre, à partir de l'analyse des citations et des références, la composition et l'évolution des publications scientifiques et au-delà, construire des réseaux des auteurs, des revues, des institutions, des pays (...) avec les différentes combinaisons possibles. Ce qui n'est rien d'autre que l'idée de la carte de la science prônée par Price (1965) et qui se base sur les "*relations structurelles du réseau de références et citations*". Concrètement, ceci revient à représenter la production scientifique sous la forme de graphe orienté avec les deux principaux éléments : les nœuds qui représentent les publications scientifiques et les arcs qui représentent les différentes relations obtenues à travers les citations. Selon Prime-Claverie (2004), les publications sont les composantes élémentaires du modèle scientifique c'est à dire les *items*. Elles sont datées et appartiennent à différentes *unités scientifiques* comme les

⁵⁹ CASE, D., HIGGINS, G. (2000). How can we investigate citation behavior ? a study of reasons for citing literature in communication. In : *Journal of the American Society for Information Science*, 51(7) : 635- 645.

auteurs, les revues, les institutions, les pays, etc. Les citations, par l'intermédiaire des références bibliographiques, relient les différents *items* ; et de manière indirecte, elles relient aussi les différentes *unités scientifiques*.

1.3.2 L'article scientifique

L'approche des citations pour aborder la production scientifique et ses impacts dans la l'organisation et l'évolution de la communauté scientifique se base naturellement sur la place qu'occupe l'article scientifique et la place et la signification que lui ont accordées différents penseurs.

Commençons par *le réductionnisme bibliométrique* que Polanco (1995) définit comme « le point de vue par lequel l'article scientifique devient un outil de définition de la science et l'on fait de la publication écrite un indicateur privilégié de l'activité scientifique, considérant que le produit final de la recherche scientifique est la publication d'un texte écrit. » Ainsi pourrait-on dire que, sont considérés comme scientifiques que ceux qui publient, et de ce fait l'article devient la chose qui matérialise l'activité scientifique. La quantité d'articles publiés fût longtemps considérée comme un indicateur pertinent de l'activité scientifique. Au 6^{ème} Congrès International d'Histoire des Sciences (Amsterdam, août 1950)⁶⁰, Price expose pour la première fois une manière d'utiliser le nombre d'articles scientifiques comme une indication quantitative de l'activité de recherche. Cette approche quantitative quant à la mesure de l'activité de la recherche sera longtemps de mise jusqu'au moment où on commence à observer une certaine dérive du côté des chercheurs qui n'utilisent plus l'article scientifique dans sa fonction première, celle de communiquer leurs savoirs, mais pour se faire reconnaître et cautionner la propriété intellectuelle de leurs travaux (Prime-Claverie, 2004). Alors Price (1963) dira que « *le degré d'utilisation semble être un meilleur test de qualité* » ; le degré et la fréquence des citations et des références reflètent même « *l'utilité des différents articles* ». Voilà ce qui sera l'hypothèse de base de l'analyse des citations de Price dans *Little Science, Big Science* (1963)

⁶⁰ Ce travail fut ensuite publié en 1951 dans la revue *Archives Internationales d'Histoire des Sciences*, vol. 14, p. 85-93.

Dès lors, la notion des citations et de son utilisation comme moyen de mesurer de manière fiable et pertinente l'activité de la science et des scientifiques sera instituée pour devenir ensuite « indispensable » en matière de métriques de la science.

1.3.3 L'analyse du graphe de citations

Il y a différentes méthodes d'analyser le graphe de citation. Nous allons seulement nous limiter ici aux notions de facteurs d'impacts et de facteurs d'influence.

Facteurs d'impacts et facteurs d'influence

« Le décompte des citations permet d'évaluer l'impact scientifique de la recherche. Le décompte des citations reçues par des revues est compilé systématiquement par Thomson ISI et vendu sous la marque de commerce *Journal Citation Reports (JCR)*. Ce produit comprend de nombreux indicateurs ayant trait aux citations reçues par les revues scientifiques et dont le *facteur d'impact* est sans doute le plus largement utilisé » (Archambault et Vignola, 2004). Ce facteur d'impact est défini comme le rapport, pour une année donnée, entre le nombre de citations des articles publiés par un périodique et le nombre d'articles publiés, le tout sur une période de deux ans. Cependant, ces facteurs d'impact présentent des limites (Pinski and Narin, 1976). D'après eux, ces dits facteurs ne tiennent pas compte, d'une part, de la longueur des articles. Ce qui fait que les articles de synthèse, plus étendus dans leur couverture et plus longs, reçoivent de ce fait plus de citations que les articles de recherche. Ensuite, ces facteurs ignorent les pratiques de citation propres aux différents domaines. Et enfin, avec l'approche des facteurs d'impacts, les citations ont la même valeur quelle que soit leur revue de provenance. En retour, ils ont présenté un nouvel indicateur, *le facteur d'influence*, pour rendre compte de l'analyse du degré de prestige des revues. Ils se sont basés sur le fait que les citations n'ont pas la même valeur, et pour cause, les revues considérées comme les plus prestigieuses reçoivent forcément plus de citations. Ce facteur d'influence est calculé à partir du poids d'influence d'une référence bibliographique et qui n'est rien d'autre que le rapport entre le nombre total de citations reçues par une revue et le nombre total de références issues de la revue. Ainsi, l'influence d'un article est égale à la somme des poids d'influence des références bibliographiques qui le citent.

II. De la bibliométrie à la wébométrie

II.1. A propos d'Internet

Le réseau Internet est né vers les années 60 au sein d'un organisme militaire américain, L'ARPA (Advanced Research Project Agency) avant de se développer dans le milieu universitaire plus tard. « *L'origine de ce projet est la construction d'un réseau informatique capable de résister à d'éventuelles attaques soviétiques, et pouvant s'auto-configurer si l'un des maillons venait à défaillir.* » (Prime-Cilverie, 2004). Le principe de base d'Internet est l'absence de structure centralisée et de « contrôle » - certains pensent pourtant qu'il existe une certaine auto-organisation ou auto-régulation du réseau (Björneborn, 2004), (nous y reviendrons) -, ce qui lui assure une expansion fulgurante et sans limite.

Le Web a été développé par Berners-Lee et ses collègues du CERN (Centre Européen de Recherche Nucléaire) à Genève en 1991 et était considéré au début comme un Intranet destiné aux chercheurs affiliés au Centre. Leur projet était de proposer un outil afin de faciliter le partage d'information entre les chercheurs du CERN, géographiquement dispersés, à travers un accès facile à des publications en ligne (Björneborn, 2004). Cette technologie a été mise gratuitement à la disposition du grand public (individus, entreprises et institutions) en 1993 (Cailliau, 1995). A partir de là, le Web va devenir un réseau gigantesque comparable à un réseau de neurones (Abraham, Ralph H., 1996). Glover et al. (2002) qualifierons le Web d'une collection de documents hétérogènes où nous retrouvons du texte, du son, de la vidéo, de l'animation (...) touchant des domaines aussi divers que le social, le culturel, l'économique, le scientifique, le politique...

II.1.1 Estimation de la taille du Web

Plusieurs études ont tenté de mesurer la taille du Web parmi lesquelles celles menées sous l'égide du NEC Research Institute en 1997 et 1999. Cette équipe, dirigée par Lawrence et Giles (1998 ; 1999), a estimé la taille du Web indexable à 320 millions de pages en 1997 et à 800 millions de pages en 1999. Leur méthode d'investigation était basée sur la combinaison de plusieurs moteurs de recherche dont AltaVista, HotBot, NorthernLight, Excite, Lycos, et Infoseek, et de recouper les réponses communes. Selon eux, le meilleur des moteurs de recherche, à l'époque Northern Light, ne pouvait couvrir plus de 16% du Web. La réunion des six plus grands moteurs de recherche ne couvrirait que 60% du Web. Mais vu que les moteurs de l'époque ne pouvaient pas indexer les pages de formats (.pdf) ou (.doc) par exemple, ce

que font maintenant les moteurs de recherche comme Google, on est tenté d'affirmer que la taille du Web était beaucoup plus grande que ne l'ont constatée les études du NEC Research Institute.

Aujourd'hui, on estime la taille du Web visible à plus de 5 milliards de pages reliées par une cinquantaine de milliards de liens hypertextes (Björneborn, 2004). D'où le constat fort parlant de Rostaing, Hervé : « *Je n'étonnerai personne en évoquant ma confusion devant l'évolution galopante d'Internet et plus particulièrement du World Wide Web* »⁶¹. L'étonnement est d'autant plus grand qu'on sait que le Web invisible, contenant les pages dynamiques et les bases de données accessibles en ligne (ex. Dialog) et que les moteurs de recherche ne peuvent pas indexer, serait 400 à 500 fois plus grande que le Web visible⁶².

La figure suivante présente l'évolution du nombre de sites Web de septembre 1995 à juillet 2003, hors sites dupliqués. Nous constatons ainsi que le nombre de sites est passé, durant cette période de 8 ans, de 18.864 sites web à plus de 42 millions sites ; ce qui nous donne une idée assez nette de l'accroissement rapide que subit le Web.

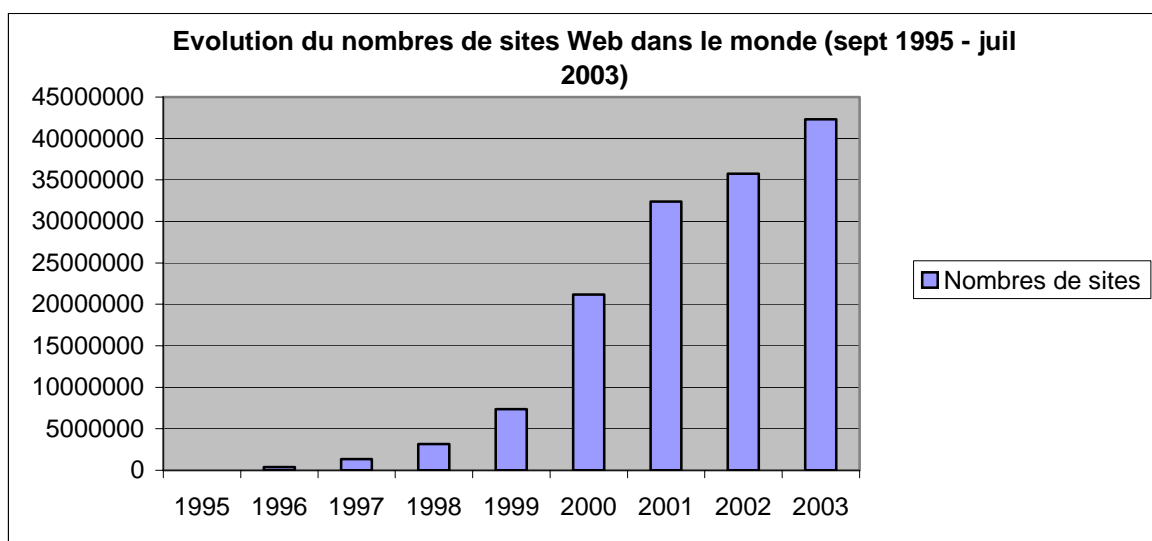


Figure 5 : Evolution du nombre de sites Web. (Sources : Le Journal du Net⁶³.)

⁶¹ ROSTAING, Hervé. Le Web et ses outils d'orientation. In : *Bulletin des Bibliothèques de France*. Paris, 2001, t. 46, n° 1, p. 68-77

⁶² How much information. [site consulté le 25/01/05].

Disponible sur l'URL : [Hhttp://www.sims.berkeley.edu/research/projects/how-much-info/internet.html](http://www.sims.berkeley.edu/research/projects/how-much-info/internet.html)

⁶³ [site visité le 25/01/04].

Disponible sur URL : [Hhttp://www.journaldunet.com/cc/03_internetmonde/intermonde_sites.shtml](http://www.journaldunet.com/cc/03_internetmonde/intermonde_sites.shtml)

II.1.2 La notion d'auto-organisation du Web

Le Web est comme un arbre constitué de domaines, de serveurs et de pages (Abraham, 1996). Avec cette structure, selon (Björneborn, 2004), le Web est devenu un système évolutif et de plus en plus complexe, contenant toute sorte d'informations, utilisées par des acteurs différents pour différentes raisons. Et comme dis plus haut, le principe de base d'Internet est l'absence de contrôle et d'organisation centralisée. Björneborn et Ingwersen (2001) caractérisent le Web de « 3D » : *distribué, diversifié et dynamique*. La distribution consiste au fait que les ressources du net sont réparties dans des millions de sites situés un peu partout dans le monde sans structure centralisée. Ces même ressources sont aussi diverses que variées et touchant toutes les activités humaines. Les rapports de recherche scientifiques, les pages de jeux, les spots publicitaires, les vitrines commerciales, les pages de propagande de toutes sortes (...), cohabitent sur le Web. Et par dynamisme, ils entendent par là le changement continu et les mutations sans arrêt que subissent les contenus des pages Web. Une page créée aujourd'hui peut disparaître du jour au lendemain ou bien changer complètement de contenu.

Avec le manque de structure centralisée et de contrôle des contenus, on est tenté de dire qu'il règne un désordre et un chaos total sur le Web. A la différence de la citation dans la littérature scientifique, *la création de liens hypertextes est moins formelle et n'est soumise à aucun contrôle* (Prime-Claverie, 2004). Et pourtant, *l'analyse du Web révèle un remarquable degré d'auto-organisation* (Björneborn, 2004). Cette auto-organisation du Web est perceptible à travers l'analyse des sujets et des centres d'intérêt des chercheurs par exemple. L'interconnexion des sites Web concernant leurs projets, leurs publications, leurs domaines et leurs institutions de recherche, est évidente. Sur ce point, l'étude de Rostaing & Boutin (1999) qui visait à cartographier la présence de la communauté des biblio-scienciométriciens sur Internet en est une parfaite illustration. Par ailleurs, la création des liens hypertextes est moins anarchique qu'on le pense. Ce processus qui consiste à se lier aux autres sites du réseau est souvent motivé par le souci de faire référence à des pages qui illustrent en quelque sorte ses propres pages, d'où l'existence un certain centre d'intérêt commun. Ce qui implique l'idée de regroupement, donc d'organisation. Nous verrons plus tard qu'il existe aussi d'autres

motivations quant à la création de liens hypertextes. Enfin, une autre manifestation de l'auto-organisation du Web se trouve dans l'apparition de plus en plus importante de sites portails et de guides spécialisés ou généraux avec comme but principal de regrouper les ressources sur un certain nombre de sujets afin de faciliter l'accès. On peut citer par exemple : SAPRISTI (Sentiers d'Accès et Pistes de Recherche d'Informations Scientifiques et Techniques sur l'Internet !)⁶⁴ élaboré par INSA de Lyon et GIRI2 (Guide des Indispensables de la Recherche sur Internet)⁶⁵ mis en place par l'Université de Laval au Canada.

II.2. La webométrie

La webométrie comme discipline spécialisée dans l'analyse des pages et sites Web (et plus précisément des liens hypertextes) est tributaire des méthodes et travaux développés dans les disciplines de métriques comme la bibliométrie, la scientométrie et l'infométrie. Cette adaptation des lois bibliométriques dans le contexte assez particulier du Web a donné naissance, et ce concrètement depuis le milieu des années 90 avec Larson (1996), à un champ d'étude très dynamique où l'on retrouve aussi bien des informaticiens, des professionnels de l'information que de mathématiciens. On peut même dire qu'elle est devenue un domaine scientifique à part entière *avec ses différentes théories à construire, des tâches à faire, des unités à définir, des méthodes à développer et des problèmes à résoudre*⁶⁶.

Par ailleurs, vu le changement qu'a introduit Internet dans la production, la diffusion et la circulation des écrits scientifiques, les professionnels de l'information, notamment, ne peuvent plus ignorer ce nouveau média. Il faut le comprendre, l'approprier à travers les outils et méthodes dont ils disposaient. « *Le Web et les autres services de l'Internet sont une aubaine pour les bibliomètres, car ils offrent de nouvelles sources d'information sur support numérique liées à l'activité scientifique (littérature grise, forums, etc.) différentes des traditionnelles bases de données d'articles* » Prime-Claverie (2004). A partir de là, plusieurs analogies entre le circuit traditionnel de la production et de l'utilisation des connaissances scientifiques et l'environnement Web vont voir le jour, et parmi lesquelles entre articles et pages Web, entre citations et hyperliens. ...

⁶⁴ Disponible sur l'URL : [Hhttp://csidoc.insa-lyon.fr/sapristi/digest.html](http://csidoc.insa-lyon.fr/sapristi/digest.html)H

⁶⁵ Disponible sur l'URL : [Hhttp://www.bibl.ulaval.ca/vitrine/giri/giri2/H](http://www.bibl.ulaval.ca/vitrine/giri/giri2/H)

⁶⁶ AGUILLO, Isidro F. (2002). "Cybermetrics : definitions and methods for an emerging discipline". *Séminaires de l'ADEST*, Paris, 14 February, 2002.

Disponible sur l'URL : [Hhttp://www.upmf-grenoble.fr/dest/seminaires/ISIDRO/Cybermetrics.ppt](http://www.upmf-grenoble.fr/dest/seminaires/ISIDRO/Cybermetrics.ppt)H

Nous reviendrons sur ces analogies, leurs applications ainsi que leurs limites, un peu plus loin. Mais commençons cette partie par définir sur le plan conceptuel et théorique ce nouveau champ de recherche.

II.2.1 Définition

Björneborn et Ingwersen (in press) définissent la webométrie comme : *“The study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web, drawing on bibliometric and informetric approaches.”*

Comme nous le voyons, cette définition englobe les aspects quantitatifs de la construction et de l'utilisation du Web. Et ainsi, la recherche en webométrie tournerait autour de quatre axes principaux. Björneborn (2004) :

- L'analyse du contenu des pages Web
- L'analyse de la structure des liens du Web
- L'analyse de l'utilisation du Web (incluant principalement les comportements de recherche des utilisateurs)
- L'analyse des technologies Web (incluant la performance des moteurs de recherche)

Par ailleurs, on voit souvent le terme cybermétrie utiliser à la place de webométrie et vice versa. Seulement pour Björneborn (2004), il existe bel et bien une nuance entre ces deux termes. Pour cela, il définit la cybermétrie comme : *“The study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Whole Internet, drawing on bibliometric and informetric approaches.”*

C'est presque la même définition que la webométrie sauf que, à la place de « *on the Web* », il met « *on the whole Internet* ». En d'autres termes, ce champ englobe les études statistiques des groupes de discussion, des mailing list et autres modes de communication sur Internet incluant bien sûr le Web. Ce qui revient à dire tout simplement que la cybermétrie englobe entière la webométrie.

Et pour résumer le tout, en tenant compte aussi de la bibliométrie, de la scientométrie et l'infométrie, il nous présente la figure suivante⁶⁷ qui montre de manière fort pertinente comment ces différentes disciplines, toutes issues des sciences de l'information, s'imbriquent les unes aux autres.

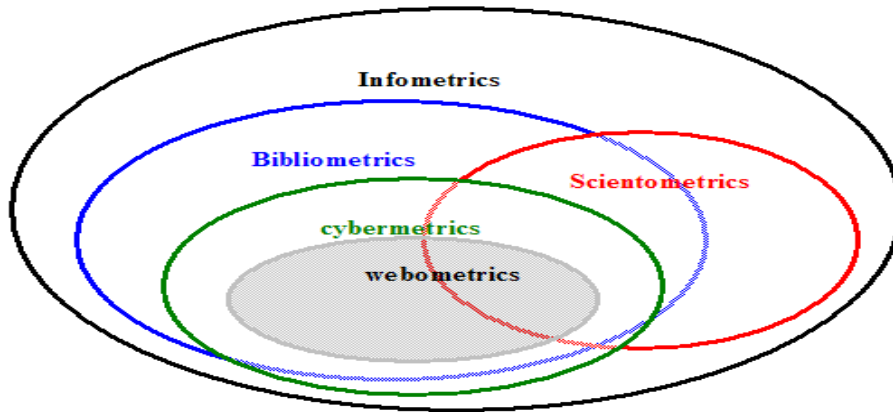


Figure 6 : Relation entre info-/biblio-/sciento-/cyber-/web-métrie (Björneborn, 2004)

II.2.2 Historique

Tout d'abord, Mike Thelwall et Han Woo Park⁶⁸ nous apprennent que le véritable intérêt des Sciences de l'information pour l'étude des liens hypertextes a commencé vers 1996 et a été principalement motivé par les analogies avec les citations des articles de périodiques. Durant cette période, plusieurs termes furent proposés pour nommer ce nouveau champ de recherche (Björneborn, 2004). A titre d'exemple on peut citer : *Netometrics* avancé par Bossy (1995) ; *Webometry* qui nous vient de Abraham (1996) ; *Internetometrics* en 1996 puis *Webometrics* en 1997 avec Almind et Ingwersen ; et enfin *Cybermetrics* coïncidant avec le début du Journal du même nom⁶⁹ en 1997 par Aguillo. Même si Chakrabarti (2002) parlera beaucoup plus tard de *Web Bibliometry*, ce sont les termes wébométrie et cybermétrie qui sont les plus utilisés.

⁶⁷ BJÖRNEBORN, Lennart. Small-world link structures across an academic web space : a library and information science approach. PhD dissertation. Copenhagen: Department of Information Studies, Royal School of Library and Information Science, 2004. p.14

⁶⁸THELWALL, M., PARK, H. W. [site consulté le 23/12/04]. Hyperlink Analyses of the World Wide Web: A Review.

Disponible sur l'URL : [Hhttp://www.ascusc.org/jcmc/vol8/issue4/park.html](http://www.ascusc.org/jcmc/vol8/issue4/park.html)

⁶⁹International Journal of Scientometrics, Infometrics and Bibliometrics.

Disponible sur l'URL : [Hhttp://www.cindoc.csic.es/cybermetrics/H](http://www.cindoc.csic.es/cybermetrics/H)

Par ailleurs, un point capital dans le développement de la wébométrie est l'émergence des moteurs de recherche commerciaux tel AltaVista qui permettait, sur une simple commande, à n'importe qui, de dénombrer les liens entre pages Web. (Park & Thelwall, 2003). Les professionnels de l'information qui ont détecté ce potentiel, n'ont pas manqué de se référer à leur propre discipline pour voir les différentes applications possibles, notamment de dresser une analogie entre articles de périodiques et documents Web, entre hyperliens et citations. Donc, selon eux le point de départ de la wébométrie est la tentative d'appliquer l'analyse des citations au contexte du Web.

II.3. Place des moteurs de recherche dans les études wébométriques

Si dans les études bibliométriques les banques de données bibliographiques (ex. ISI Thomson) et autres bases dédiées à la compilation des écrits scientifiques fournissent les corpus et les échantillons de traitement, en wébométrie c'est les moteurs de recherche qui jouent, à quelques différences près, ce rôle. Mais qu'est-ce qu'un moteur de recherche ?

Un moteur de recherche est un programme qui indexe automatiquement les pages Web. En suivant les hyperliens, il repère et collecte les pages, extrait tous les mots (sauf les mots vides) contenus dans ces pages et en fait une base de données. Il lie ainsi, à travers un système d'appariement, cette base de données ainsi constituée et les utilisateurs. Mais répondent-ils vraiment aux attentes des wébomètres ?

II.3.1 Utilisation et limites des moteurs

Les modes de recherches avancées des moteurs permettent aux wébomètres des opérations booléennes plus complexes, donc des recherches plus ciblées. Citons par exemple les opérations : *link, domain, site, host, title, ...* L'utilisation des moteurs de recherche de première génération comme Alta Vista, Nothern Light, HotBot en wébométrie ont montré très vite les limites de ces outils.

Et même si les algorithmes de ces moteurs sont devenus de plus en plus développés, comme abordé plus haut, leur couverture du Web est très limitée (Lawrence et Giles, 1998). D'autres problèmes concernent le flou qui règne dans la fréquence des mises à jour, des règles d'indexation, des algorithmes de classement. Sur ce dernier point, notons l'innovation du moteur Google, (Brin & Page, 1998), avec son algorithme *Page Rank* qui prend en compte la dimension structurelle du Web et classe ainsi les pages en fonction du nombre de liens qui

pointent vers elles. Ce qui n'est rien d'autre que l'application du facteur d'influence adapté au graphe du Web (voir page 41).

Par ailleurs, Rostaing dénote d'autres faiblesses et erreurs des moteurs de recherche comme : *des pages supprimées dans les sites mais maintenues dans l'index, des pages modifiées dans les sites et toujours caractérisées par les mots de l'ancienne version dans l'index, des pages de grandes tailles indexées uniquement avec un ensemble restreint de premiers mots, la disparition de pages de l'index alors qu'elles sont toujours présentes dans les sites, la disparition de mots caractérisant une page sans que la page ait été modifiée*⁷⁰.

Enfin, l'utilisation des moteurs comporte aussi d'autres problèmes. En plus de la limitation causée par leur incapacité à couvrir la totalité du Web, il y a une autre limitation qui est cette fois-ci volontaire et relève de la part des concepteurs de ne pas dévoiler la totalité de leurs informations (Prime-Claverie, 2004). Par exemple, avec une recherche sur Google avec la fonction *site*, il est impossible d'extraire plus de 300 références quel que soit le nombre de résultat trouvé par le moteur.

Ainsi devenons-nous faire avec ces limites et nous contenter de ces outils au risque de produire des travaux de qualité moindre ? Ou bien, devenons-nous développer des outils alternatifs mieux adaptés au domaine des sciences de l'information et qui seront à même de répondre aux attentes des webomètres ?

II.3.2 Quelques réponses de professionnels de l'information

Cette partie a pour base et pour point de départ l'appel de Bar-Ilan⁷¹ à la communauté des sciences de l'information à avoir ses propres moteurs (crawler), accessibles à tous et qui permettront des méthodes de collecte de données fiables et transparentes.

S'il y a un groupe de recherche qui a vraiment œuvré dans ce sens, c'est bien l'équipe de Mike Thelwall : *The Statistical Cybermetrics Research Group*⁷² de l'Université de Wolverhampton en Angleterre. Connaissant la difficulté à bien parcourir le Web pour

⁷⁰ Ibid

⁷¹ BAR-ILAN, J. (2001). "How much information the search engines disclose on links to a web page? – A case study of the 'Cybermetrics' home page." In : *Proceedings of the 8th international Conference on Scientometrics and Informetrics*, 1, 63-73.

⁷² [Hhttp://cybermetrics.wlv.ac.uk/H](http://cybermetrics.wlv.ac.uk/H)

constituer un corpus de travail, ils ont développé et mis à la disposition des professionnels, gratuitement, des bases de données⁷³ des structures des liens hypertextes de plusieurs universités : Royaume-Uni, Nouvelle Zélande, Australie, Chine, Taïwan, ... Pourquoi les sites universitaires ? Pour Thelwall⁷⁴, il existe deux raisons pour cela : d'une part, concernant l'utilisation d'Internet, le secteur académique est plus mature que les autres secteurs ; d'autre part, les sites Web des universités permettent une comparaison très nette avec les articles des travaux universitaires. Ce qui explique aussi par ailleurs pourquoi la plupart des études wébométriques et cybermétriques concerne le milieu universitaire.

En plus de ces bases de données, l'équipe de Wolverhampton a mis aussi en accès libre, toujours sur son site, un crawler de liens hypertextes *Soscibot*. Il permet entre autre, de parcourir et d'identifier les liens entrants et les sortants d'un site Web donné. Nous reparlerons de cet outil dans la prochaine partie.

De telles tentatives et initiatives montrent à la fois la jeunesse mais aussi le dynamisme de cette nouvelle discipline qui s'affirme de plus en plus. L'intégration et la prise en compte par ces outils des autres secteurs seraient une excellente chose. Car le but de tout cela est d'arriver à avoir des données fiables et pertinentes pour procéder à de bonnes analyses.

II.4. Analyse du graphe du Web

L'un des qualificatifs que l'on donne le plus souvent à Internet est : le réseau des réseaux. Ce qui implique naturellement l'idée de représentation, de graphe, de liens, de relations, d'interconnexion... « *Le Web peut être modélisé comme un graphe mathématique en considérant ses pages comme des nœuds et comme arcs, les liens hypertextes.* »⁷⁵. Et pour Ingwersen et Björneborn (2001), la théorie des graphes est un excellent outil pour comprendre la structure des liens du Web. De manière très particulière, ces liens hypertextes représentent une importance de premier ordre en ce qu'ils déterminent même la structure mais aussi l'expansion et la taille de plus en plus grande du Web. Car, grâce à ces liens, créer sa page Web, s'ancrer aux autres sites et s'inviter ainsi au réseau global devient de plus en plus chose aisée,

⁷³ [Hhttp://cybermetrics.wlv.ac.uk/database](http://cybermetrics.wlv.ac.uk/database)H

⁷⁴ THELWALL, Mike. [site visité le 23/12/04]. A Free Database of University Web Links: Data Collection Issues. In : *Cybermetrics. Issues Contents*: Vol. 6/7 (2002/3) : Paper 2, 11p.

Disponible aussi sur l'URL : [Hhttp://www.cindoc.csic.es/pruebas/v6i1p2.htm](http://www.cindoc.csic.es/pruebas/v6i1p2.htm)H

⁷⁵ THELWALL, M., WILKINSON, D. (2002). Graph Structure in Three National Academic Webs : Power Laws with Anomalies. In : *Journal of the American Society for Information Science and Technology*, 54(8), 706-712. Disponible aussi l'URL : [Hhttp://cybermetrics.wlv.ac.uk](http://cybermetrics.wlv.ac.uk)H

d'où la croissance exponentielle du Web (Larson, 1996). Par ailleurs, "*The study of the structure of this graph is useful because of the importance of hyperlinks for search engine web crawlers and in information science web link research*". (Björneborn, 2001). C'est pourquoi Han Woo Parker⁷⁶ dit que : « *L'élément structurel de base d'Internet est le lien hypertexte* ».

Mais avant d'entrer dans le vif du sujet et de montrer le caractère spécifique par rapport à la théorie des graphes en science sociale ou en bibliométrie, nous commencerons par quelques définitions opérationnelles sur le Web mais aussi sur les différents types de liens hypertextes.

II.4.1 Quelques définitions opérationnelles

Le Web, l'environnement Internet en général, dispose de ses propres termes et concepts qui permettent de bien le décrire et de le différencier de tout autre environnement. Un éclaircissement sur ces termes du point de vue conceptuel ne peut qu'être une chose nécessaire et même incontournable pour notre appréhension et notre compréhension, d'une part. D'autre part, cela nous permettra de bien cerner la relation qu'entretiennent ces différents éléments et comment ils sont structurés.

❖ Quelques termes du web

Les termes les plus « importants » du Web et que nous allons fréquemment utiliser dans cette étude sont : site Web, page Web, serveur Web, nom de domaine, URL :

- **Un site Web** est un emplacement donné par un nom de domaine contenant une ou plusieurs pages Web, reliées par des liens hypertextes ou des images ancrées. Ces sites sont créés et maintenus par un individu, une compagnie ou une organisation⁷⁷.
- Si un site peut être conçu comme un terme Web, et représentant un document Web, **le serveur Web** est quant à lui un terme d'Internet représentant une ou plusieurs machines ou ordinateurs (Björneborn, 2004). Pour lui, cette distinction

⁷⁶ PARKER, Han Woo. (2003). Hyperlink Network Analysis: A New Method for the Study of Social Structure on the Web. In : *Connections*, 25(1): 49-61

⁷⁷ [Hwww.webopedia.com/TERM/w/web_site.html](http://www.webopedia.com/TERM/w/web_site.html) . [site visité le 05/01/05]

conceptuelle est essentielle car le Web et Internet sont deux entités différentes. Si le Web est un réseau de documents reliés par des liens hypertextes, Internet est un réseau de machines reliées par des câbles et des routeurs.

- **Un nom de domaine** fonctionne comme un système d'adressage et d'identificateur avec un nom alphanumérique utilisé pour identifier une ou plusieurs adresses IP. Vu que Internet est basé sur l'adressage numérique (IP) et non sur les noms de domaine, chaque serveur Web a besoin d'un DNS (Domain Name Server) pour traduire les noms de domaine en adresse IP. Un nom de domaine basique est composé de trois segments : www.xxx.yy. Le dernier segment (yy), le Top Level Domain (TLD), peut désigner le code de domaine d'un pays (ex. .fr pour la France, .sn pour le Sénégal) ou le type de site : com, edu, gov, coop, ...
- **L'URL** (Uniform Resource Locators) est un système standardisé d'attribution des adresses sur Internet. « Les URL identifient les ressources sur le Web : documents, images, fichiers téléchargeables, services, boîtes de messagerie électronique et autres ressources ... » (World Wide Web Consortium, 2002). Par extension, l'URL désigne aussi l'adresse d'un site ou d'une page Web. L'adresse URL complète est composée :

- 1 - du type de protocole (http, ftp ou gopher)
- 2 - du nom du serveur (ou nom de domaine)
- 3 - de l'emplacement exact du fichier

Exemple : <http://www.sonatel.sn/abonnements.htm>

❖ Les différents types d'hyperliens

S'il est acquis que le Web est aujourd'hui considéré comme un graphe avec comme nœud une page web par exemple et les hyperliens comme arcs, il n'en demeure pas moins que la nature de ces derniers (les hyperliens) ne sont pas toujours nette et clairement définie. Par exemple, on divise souvent les hyperliens en deux types : liens internes et liens externes. La définition ou la limitation de ces liens externes pose problème puisqu'il peut s'agir soit de liens sortant du site concerné vers d'autres sites ; soit des liens venant d'autres sites vers le site concerné. (Björneborn, 2004). Donc, tout cela mérite qu'on essaye d'y voir un peu plus

clair, et c'est ce que nous allons faire en nous référant principalement aux notions développées, à travers un graphique, par Lennart Björneborn dans sa thèse.⁷⁸

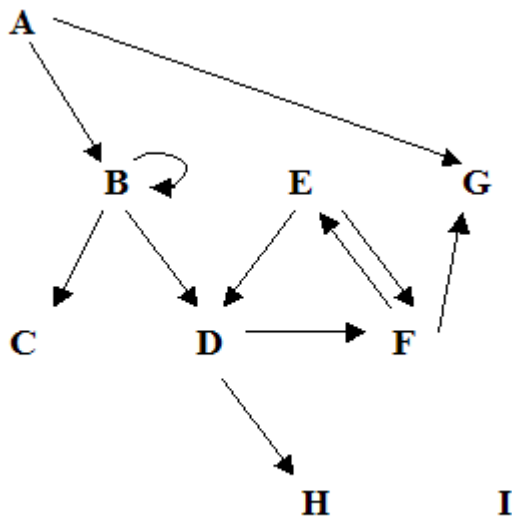


Figure 7 : Terminologie de base des liens wébométriques (Björneborn, 2004)

Les lettres (A, B, C, D, E, F, G, H et I) désignent les nœuds et peuvent être des pages Web, des sites Web, des répertoires... Et les flèches sont les liens hypertextes qui relient ces différents éléments. Pour la compréhension des relations, traduisons les termes *Inlink* et *Outlink* respectivement par *lien entrant* et *lien sortant*. Ainsi, on a les relations suivantes :

- B has an inlink from A; B is inlinked; A is inlinking; A is an in-neighbor of B
- B has an outlink to C; B is outlinking; C is outlinked; C is an out-neighbor of B
- B has a selflink; B is selflinking
- A has no inlinks; A is non-linked
- C has no outlinks; C is non-linking
- I has neither in- nor outlinks; I is isolated
- E and F have reciprocal links; E and F are reciprocally linked
- D, E and F have in- or outlinks connecting each other; they are triadically interlinked
- A has a transversal outlink to G: functioning as a shortcut
- H is reachable from A by a directed link path

⁷⁸ BJÖRNEBORN, Lennart. Small-world link structures across an academic web space : a library and information science approach. PhD dissertation. Copenhagen: Department of Information Studies, Royal School of Library and Information Science, 2004. xxxvi, 399 p.

- C and D are co-linked by B; C and D have co-inlinks
- B and E are co-linking to D; B and E have co-outlinks
- Co-inlinks and co-outlinks are both cases of co-links

II.4.2 Citation et « Sitation »

Le terme *sitation*, désignant la relation entre deux sites Web a été prononcé pour la première fois en 1996 par McKiernan⁷⁹ et a été utilisé par Aguillo lors de la conférence de 4S/EASST à Bielefeld en octobre 1996 (Rousseau, 1997). Ronald Rousseau⁸⁰ a été sans doute le chercheur qui a véritablement popularisé ce concept (Thelwall, 2003). Cette notion, comme montré plus haut, s'inscrit dans une tentative de faire une analogie entre le Web et les publications scientifiques. Selon Rousseau, étudier la notion de *sitation* est le même, sur le plan conceptuel, qu'étudier la citation entre articles de périodique. Cependant, il y a une certaine différence dans les significations. A la différence de la citation, la *sitation* est rarement utilisée pour *argumenter, comparer ou présenter des idées* (Chu, 2004). Généralement, son objectif est de faire référence à un site intéressant. Elle cible soit une page Web soit le contenu d'un site entier, alors que la citation est beaucoup plus précise, en ce sens qu'elle peut se porter uniquement sur une phrase ou un paragraphe.

Mais essayons de comprendre les motivations qui font qu'un site Web « *site* » un autre site Web.

❖ « Sitations » : motivations

Notons qu'il n'existe pas de règles quant à la création d'hyperliens. Il n'y a pas de règles codifiées et reconnues et à partir desquelles les motivations de créations de liens hypertextes se justifient comme c'est le cas dans les publications scientifiques (Ingwersen & Björneborn, 2001). Cette irrégularité et ce désordre sont décrits par Mike Thelwall (2003) : « *Web links represent both anarchy and order* ». Selon lui, l'ordre est perçu, par exemple, à travers les moteurs de recherche comme Google ou Alta Vista qui, justement, utilisent avec succès la structure des liens hypertextes pour optimiser les résultats de recherche.

⁷⁹ MCKIERNAN, G. CitedSites(sm): Citation Indexing of Web resources.
Disponible sur l'URL : [Hhttp://www.public.iastate.edu/~CYBERSTACKS/Cited.htm](http://www.public.iastate.edu/~CYBERSTACKS/Cited.htm)

⁸⁰ ROUSSEAU, Donald. [site visité le 06/01/05]. Sitations ; an exploratory study.
Disponible sur l'URL : [Hhttp://www.cindoc.csic.es/cybermetrics/articles/v1i1pl.html](http://www.cindoc.csic.es/cybermetrics/articles/v1i1pl.html)

Comprendre la structure des liens passe incontestablement par la compréhension des différentes raisons qui poussent un site Web à « *siter* » un autre site. C'est ce que Thelwall a essayé de faire, dans un article précurseur⁸¹, dans le cadre universitaire.

Selon lui, il faut d'abord commencer par faire une différenciation entre liens intra-sites qui relient des pages hébergées sur le même site et liens inter-sites qui relient des pages hébergées sur des sites différents.

Sa base de travail est constituée des liens hypertextes de 111 universités britanniques. Sur un total de 19.438 liens, il en a choisi 100 au hasard comme corpus pour cette étude. Il est arrivé ainsi à dégager quatre catégories de motivations :

1. **General navigational links** (les liens de navigations générales)
2. **Ownership links** (les liens de propriété)
3. **Social links** (les liens sociaux)
4. **Gratuitous links** (les liens gratuits)

- **Les liens de navigations générales**

Un lien est décrit comme étant un lien de navigation générale si la motivation première de sa création est de constituer un point de départ afin de permettre aux visiteurs d'accéder à d'autres informations - contenues dans d'autres sites - qui ne rentrent pas forcément dans les thèmes du site en question. Ces liens jouent en quelque sorte le rôle des *renvois d'orientation* qu'on retrouve en documentation, seulement à la différence des dits renvois, il n'existe pas de relation de sens, pas de connexion cognitive entre la page source et la page cible.

- **Les liens de propriété**

Ces liens *permettent de revendiquer la propriété intellectuelle d'un document*. A l'heure des travaux collaboratifs et des projets co-dirigés, ces liens apparaissent comme manifestant une appartenance commune entre les différents partenaires. En général, les informations et données relatives aux projets ou travaux partagés par le « *collaboratoire* » sont hébergées sur

⁸¹ THELWALL, Mike. [site visité le 23/12/04]. What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. In : *Information Research*, Vol. 8 No. 3, April 2003. Disponible aussi sur l'URL : [Hhttp://informationr.net/ir/8-3/paper151.html](http://informationr.net/ir/8-3/paper151.html)H

le site de l'un des participant ou sur un serveur commun. Sur les sites des différents membres, on trouve souvent un menu faisant référence aux projets communs et renvoyant aux différents partenaires. Selon Thelwall, ces liens peuvent aussi être considérés comme des *remerciements implicites*.

- **Les liens sociaux**

De manière générale, ce sont des liens vers des collaborateurs et partenaires. D'une manière plus précise, ce sont des liens créés dans l'optique de renforcer un lien ou une relation sociale. Pour Thelwall, ces liens peuvent être perçus comme un *compliment implicite*. On reconnaît l'importance d'un site, et de ce fait, on juge utile de créer un lien vers lui. C'est une catégorie de liens très intéressante à étudier mais dont les motivations sont difficiles à déterminer.

- **Les liens gratuits**

Ces liens sont créés sans aucune motivation de communication particulière, et de ce fait, on ne s'attend pas à ce qu'ils jouent un quelconque rôle. Par exemple, ce sont les liens qui font référence aux universités où l'on a fait ses études, aux entreprises où l'on a pu travailler...

Voilà les quatre catégories qui regroupent les différentes raisons qui peuvent pousser à « *siter* » une page ou un site Web. Mais selon (Prime-Claverie, 2004), cette catégorisation manque un peu d'exhaustivité à cause notamment du contexte d'étude ou du cadre d'investigation dans lequel ces motivations ont été dégagées. Il s'agit du milieu universitaire. Selon elle, les pages d'accueil des universités (qui composent le corpus de Thelwall) comportent rarement des informations de fond, ce qui fait qu'il y a ni liens cognitifs, ni liens thématiques dans l'expérience sus présentée.

Ainsi propose t-elle de compléter la liste de Thelwall par :

- *Les liens de navigation thématique, permettant la navigation entre pages de même thème,*
- *Et les liens cognitifs, qui pointent vers des pages évoquant ou argumentant les idées de la page initiale.*

Enfin, elle propose d'inclure dans les *liens gratuits*, les liens de publicité, qui ne rapportent rien en terme de sémantique ou de cognition mais qui comptent beaucoup financièrement.

❖ Limites de l'analogie

Comme nous l'avons vu depuis le début de cette deuxième partie, la naissance et le développement de la webométrie ont pour base, principalement, l'application des méthodes biblio-sciento-métriques et plus particulièrement l'analogie entre articles scientifiques et pages Web. Cette tentative d'analogie présente pas mal de limites. Prime-Claverie, Beigbeder et Lafouge (2002) nous en donnent quelques-unes :

- Une différence majeure entre un article scientifique et une page web réside dans la volatilité et la possibilité de mise à jour de la page web. Rien ne certifie le changement ou même la disparition pure et simple d'une page *sitée* par une tierce page. Ce qui pose naturellement un problème de pertinence et de fiabilité des *sitations*.
- Comme nous le savons, la relation de citation entre deux auteurs n'est jamais réciproque, puisqu'on cite une référence qui est antérieure à l'article qu'on va publier. Alors que dans l'environnement Web, il est tout à fait possible que deux pages Web se *sitent* mutuellement. Ainsi, le caractère unidirectionnel du graphe de citations disparaît pour le Web.
- Le phénomène de duplication est très fréquent sur le Web. Cette procédure a pour objectif de permettre un plus rapide accès aux ressources. *Certains serveurs très volumineux et souvent consultés évitent les encombrements en proposant plusieurs copies de leurs sites en différents points de la planète. On parle alors de sites miroirs.* Cette pratique a pour conséquence de générer aussi la multiplication des liens hypertextes, ce qui va fortement biaiser l'analyse du graphe du Web.
- Comme nous l'avons vu dans la précédente section, les motivations de *sitation* sont multiples et diverses. Les liens de navigation et les liens gratuits ou de publicité très fréquents sur le Web ne peuvent pas être placés au même titre qu'une citation puisqu'ils sont dépourvus de sens et de signification.

II.4.3 Le degré de connectivité du Web

Dans un article assez répandu, « Diameter of the World-Wide Web »⁸², Albert et al. (1999) ont tenté de calculer le diamètre du Web, c'est à dire la chaîne la plus longue entre deux pages Web. Au moment où la taille du Web était estimée à 800 millions de pages (1999), ils ont pu arriver à la conclusion suivante : en choisissant par hasard deux pages Web, on peut passer de l'une à l'autre en 19 clics en moyenne. En d'autres termes, ils considéraient le Web comme un univers de faible diamètre et fortement interconnecté. Cette notion de « *small world* » (petit monde) importée de l'analyse réseau en science sociale pour caractériser le Web, sera ultérieurement contestée par Border et al. (2000) à travers une étude restée référence. Ils ont constitué un corpus de 200 millions de pages par le biais du moteur de recherche Alta Vista. La figure suivante montre des aspects très intéressants de la connectivité du Web assez loin des conclusions de Albert et al.

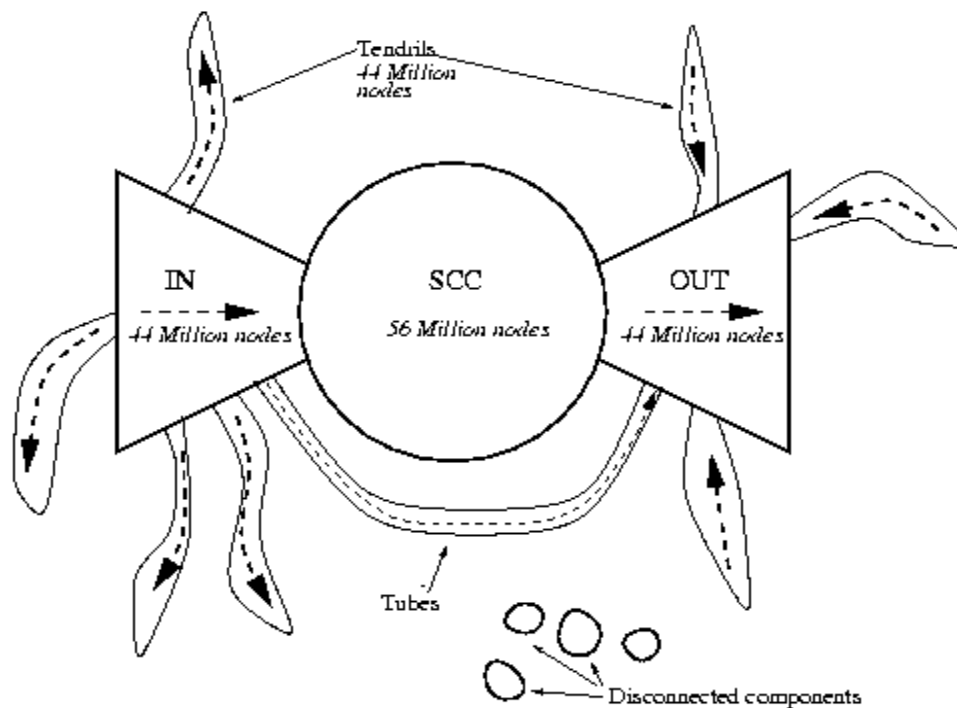


Figure 8 : Connectivité du Web (Broder et al., 2000)

⁸² ALBERT, R., JEONG, H., BARBARASI, A.-L.. Diameter of the World Wide Web. In : *Nature* 401:130-131, Sep 1999.

Leur principale découverte était que, le corpus ainsi constitué pouvait être divisé en 5 grands ensembles, chacun avec ses caractéristiques et son degré d'orientation et de connexion : Strongly Connected Component (SCC), IN, OUT, Tendrils et Disconnected.

Tout d'abord, il y a le (SCC), *Strongly Connected Component* qui peut être traduit par Composantes Fortement Connexes (Prime-Claverie, 2004). Cet ensemble qui est au fait le cœur de tous les ensembles est constitué de 56 millions de pages sur les 200 millions composant le corpus. C'est le seul ensemble où toutes les pages sont reliées les unes aux autres par un chemin. Son diamètre est estimé à 28 liens. Le concept de « petit monde » peut s'appliquer à cet ensemble.

Ensuite, nous avons les ensembles *OUT* et *IN* contenant chacun 44 millions de pages. Si les pages de l'ensemble *OUT* ne peuvent être atteintes qu'à partir du *SCC*, celles de l'ensemble *IN* peuvent atteindre les pages du *SCC* directement. Ce qui revient aussi à dire que une recherche de liens lancée à partir de l'ensemble *IN* contiendra les pages de l'ensemble *SCC* plus celles de l'ensemble *OUT*.

Nous avons aussi les *Tendrils*, qui contiennent 44 millions de pages ne pouvant ni atteindre l'ensemble *SCC* ni être atteintes à partir de celui-ci.

Enfin, il reste l'ensemble *Disconnected* contenant 16 millions de pages. Et comme son nom l'indique, il n'est lié à aucun des quatre ensembles sus-cités et est complètement déconnecté.

Par ailleurs, ils ont aussi émis l'idée d'un possible passage ou liaison d'une petite partie de l'ensemble *IN* vers une petite partie de l'ensemble *OUT* sans passer par le cœur, formant ainsi un *Tube*.

Cette découverte montre que le Web est loin d'avoir l'aspect d'un « petit monde » où il y aurait un fort degré d'interconnexion. Les auteurs ont pu estimer le diamètre du graphe (dressé à partir des 800 millions de pages extraites), à 500. Ils ont aussi montré que, en choisissant au hasard deux pages, la probabilité pour qu'il existe un chemin entre elles est de 24%. S'il s'agit d'un chemin direct, sa longueur moyenne est estimée à 16. Dans le cas d'un chemin indirect, c'est à dire que les liens entre ces deux pages vont dans les deux sens, la longueur du chemin est estimée à 6.

II.4.4 La notion de Web Impact Factor (WIF)

Le Web Impact Factor est un outil quantitatif pour classer, catégoriser et comparer des sites Web, des pages web et des noms de domaine. Essentiellement, il évalue l'impact d'un site Web à travers le dénombrement des liens entrants c'est à dire le nombre de liens qui pointent vers le site et de liens sortants c'est à dire des liens qui partent du site vers d'autres sites. Comme c'est le cas de plusieurs notions du champ de la webométrie, ce concept est basé aussi sur l'analogie entre citations et liens hypertextes et s'inspire de ce fait du *Journal Impact Factor* de l'ISI (voir page 41).

Cette notion a été introduite en 1998 par Ingwersen même si certains pensent que l'étude des facteurs d'impact sur Internet a été abordée pour la première fois par Rodriguez Gairin en 1997 dans le Journal Espagnol de la Documentation (Björneborn, 2004). Seulement, il n'a pas été aussi influent qu'Ingwersen. Ce dernier détermine trois types de Web Impact Factor : interne, externe et global. Le WIF interne est égal au rapport entre le nombre de liens entrant dans un site ou un domaine et le nombre de pages web contenues dans le site ou le domaine en question. Le WIF externe se calcule par le nombre de liens sortant d'un site web ou d'un domaine divisé par le nombre de pages web contenues dans le site. Enfin, pour le WIF global, nous avons toujours le même dénominateur (le nombre de pages contenues dans le site ou le domaine en question) mais le numérateur est égal à l'ensemble des liens externes (entrants comme sortants).

Noruzi (2004)⁸³ nous énumère quelques avantages et limites de l'approche WIF parmi lesquels :

❖ Avantages

- ✓ Il permet d'évaluer l'importance relative d'un site web en le comparant notamment aux autres sites dans un champ ou dans un nom de domaine d'un pays ;

⁸³ NORUZI, Alireza. [site visité le 09/02/05]. The Web Impact Factor: Advantages and Disadvantages. Disponible sur l'URL : [Hhttp://cybermetrics.wlv.ac.uk/AoIRASIST/Noruzi_full.htm](http://cybermetrics.wlv.ac.uk/AoIRASIST/Noruzi_full.htm)

- ✓ Il permet de faire ressortir la visibilité et la popularité d'un site Web, mais aussi la visibilité d'une compagnie, d'une organisation ou d'un pays dans la toile mondiale ;
- ✓ Le WIF et les liens externes sont utilisés dans les systèmes PageRank par certains moteurs de recherche comme Google pour classer notamment les résultats de recherche ;
- ✓ Il permet de mesurer le succès et l'influence globale d'un site Web ou d'un domaine ;
- ✓ Etc.

❖ **Limites**

- ✓ Le principal inconvénient du WIF est qu'il est influencé pour une grande partie par la couverture des moteurs de recherche. Aussi bien pour le nombre de liens entrants et sortants que pour le nombre de pages contenues dans le site en question, cela dépend du degré de couverture du moteur de recherche utilisé. Et quand on sait que, théoriquement, la combinaison des meilleurs moteurs de recherche ne couvre que près de 60% du Web global (Lawrence & Giles, 1998), cela constitue une réelle limite pour le WIF ;
- ✓ Il y a un biais introduit par les langues de publications sur le net. Les pages Web développées en langue anglaise (qui domine le Web), auront forcément un WIF plus important que les autres ;
- ✓ Il n'y a pas de différence entre d'une part, le site Web A qui contient 10 pages Web et génère 10 liens et d'autre part le site Web B qui contient 100 pages et génère 100 liens ;
- ✓ Le WIF d'un site Web est déterminé généralement sans tenir compte de la qualité scientifique des pages contenues ;
- ✓ Etc.