Ecole Nationale Supérieure
des Sciences de l'Information
et des Bibliothèques

Diplôme Professionnel Supérieur en Sciences de l'Information et des Bibliothèques

Rapport de Recherche Bibliographique

Le Dynamisme du World Wide Web: Taille, Croissance, Visibilité Distribution et Accessibilité de l'Information

Rehab OUF

sous la direction de
M. Jean-Pierre LARDY
Maître de Conférence – Co-directeur de l'URFIST de Lyon
Université Claude Bernard Lyon 1

2000-2001

Remerciements

Mes remerciements vont aux responsables de la bibliothèque de l'ENSSIB, qui ont constitué une excellente collection de périodiques en science de l'information.

Mes remerciements vont aussi à SOAD ODDEH, doctorante à l'ENSSIB, qui m'a fourni un lien très important qui m'a servi dans la rédaction.

Toute ma reconnaissance à Monsieur Jean-Pierre LARDY, mon commanditaire, qui m'a proposé un sujet très passionnant et dont la valeur ajoutée, pour moi,, est inestimable.

"When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of science."

Lord Kelvin

Le Dynamisme du World Wide Web: Taille, Croissance, Visibilité Distribution et Accessibilité de l'Information

Rehab OUF

RESUME

Dans ce vaste paysage électronique chaotique et en pleine mutation qu'est le World Wide Web, il n'existe pas à l'heure actuelle une loi qui gère le contenu qui y est édité: l'accès à un serveur Web est la seule condition pour publier de l'information sur le réseau. En plus, le contenu du Web est sujet à de perpétuels changements: par l'ajout de nouveaux matériaux, et le remplacement, la modification ou la suppression de matériaux existant. Ceci dit, estimer la taille du Web d'une manière précise, est un défi à relever.

MOTS-CLES[♣]

Taille du Web
Croissance du Web
Web Visible
Web Invisible

ABSTRACT

The World Wide Web is an unstructured, dynamic collection of electronic information. There is no selection policy governing its content: access to a Web server is the only prerequisite for publishing information on the Web. In addition, the content of the Web is subject to continuous change, as new material is added and existing material is replaced, updated, or simply disappears. Given these complications, estimating the size of the Web in a precise way is a challenging task.

KEYWORDS

Web size
Web growth
Visible Web
Invisible Web

[♣] Le terme « mots-clés » est plus approprié que celui de « descripteurs » pour à sujet de recherche, car étant donné la nouveauté de celui-ci, aucun descripteur ne lui a été encore attribué.

TABLE DES MATIERES

I- METHODOLOGIE

I- METHODOLOGIE

Vu la spécificité et la nouveauté du sujet, et l'absence totale de livres qui en traitent, la recherche a été effectuée principalement sur le Web, puis elle a été complétée par l'interrogation de certaines bases de données du serveur DIALOG, et par les périodiques en papier.

I-1 La Recherche sur le Web

La recherche sur le Web a jalonné toutes les étapes de la présente étude. Elle s'est étalée sur toute la période de la recherche, elle s'est effectuée à plusieurs niveaux par le biais d'outils très diversifiés. Dans l'ensemble, on peut dire que c'était la recherche fondamentale du travail, et qu'elle s'est caractérisée par son évolutivité.

I-1-1 Le Web comme premier recours

Au tout début la recherche a été exclusivement menée sur le Web, non seulement parce que le sujet s'y rattache tout court, mais aussi c'était une nécessité dictée par plusieurs contraintes:

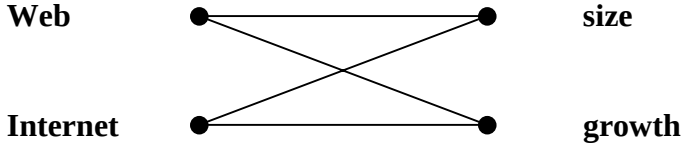
- l'absence totale de monographies consacrées exclusivement au sujet (donc pas de recherche par le titre)
- la nouveauté du sujet qui a fait qu'aucun descripteur ne lui a été encore attribué
- ce dernier point a interdit la découverte d'ouvrages dont un ou plusieurs chapitres ont été consacrés au sujet (donc pas de recherche par sujet)

Comme tout utilisateur obsédé par une question, qui est encore dans une période de tâtonnement et d'incertitude, le Web est toujours le premier recours:

- il fournit l'information de base pour un sujet donné
- il permet de découvrir le sujet, de se familiariser avec lui et de le bien cerner
- il sert à définir les mots-clés et à discerner les termes qui génèrent du bruit ou qui renvoient à des documents non pertinents

1.1.1.1. Cerner le sujet et collecter des mots-clés

Dans un premier temps la recherche s'est faite avec des mots-clés de notre choix, en combinant les termes qui expriment la "taille" et la "croissance" avec les variantes du mot "Web":



Ces différentes combinaisons insérée, chacune, entre guillemets « » ont constitué les requêtes d'interrogation des différents moteurs de recherche et métamoteurs sans

distinction, car on cherche ici des termes exacts, ça n'a rien à faire avec la syntaxe d'interrogation de chaque outil.

Ce moyen semble le plus adapté à des outils qui indexent et recherchent du "texte intégral", car il réduit au minimum le bruit en n'affichant que des documents qui contiennent l'expression exacte.

Pour les moteurs de recherche, les 60 premiers résultats dans le listing de chacun ont été examinés, pour les métamoteurs les 20 premiers résultats.

- Les principaux moteurs interrogés sont:

  Google, Altavista, Go, HotBot, Snap, Lycos, Microsoft, Northerlight (Special Collection)

- Les métamoteurs:

  Copernic, LexiBot (clients), Inxquick (enligne)

Résultats de la Recherche

1) Le terme "Web" est plus approprié au sujet que le terme "Internet" qui génère trop de bruit.
2) Les documents non pertinents abordaient la "croissance du Web" des points de vue:
   a) de la taille et la croissance en nombre de connectés
   b) de l'étendue et l'expansion du réseau
   c) "la population" ou la "démographie" de l'Internet c.a.d. les utilisateurs en nombre, en catégories, en tranches d'âge et en genre
   d) le trafic sur le réseau et les notions de télécommunication et de paquets d'information
   e) la taille et/ou la croissance du commerce électronique ou du ebusiness
   f) "usability" et la mesure du trafic sur certains sites
3) des références pertinentes mais anciennes (une date antérieure à 1997) ont été obtenues
4) Cette première récolte a permis la définition d'autres mots-clés dont:
   a) Web estimates/estimations
   b) Internet/Web Statistics
   c) Indexable Web
   d) Visible Web
   e) Invisible Web
5) Les 3 derniers termes ont été détectés chaque fois qu'un document pertinent (parlant du volume de l'information sur le Web) était retrouvé.

I-1-2 Recherche par les mots-clés les plus appropriés

Le sujet maintenant cerné, les mots-clés les plus pertinents discernés, on a lancé une nouvelle recherche sur le Web comme la précédente sans troncatures, ni opérateurs booléens, ni trop de complications mais tout simplement en utilisant les 3 expressions suivantes qui semblent comme les maîtres-mots de la recherche:

"Indexable Web"	**"Visible Web"**	**"Invisible Web"**

Toutefois un affinage a été fait sur la base de la date (à partir de 1998)

Grâce à l'utilisation de ces mots-clés, le degré de pertinence a considérablement augmenté. Cependant, ces références ont été filtrées en les soumettant à une grille d'évaluation. Ce traitement a été généralement effectué au moment même de la recherche, certains document trop longs ou qui présentent certaines difficultés ont été imprimés ou sauvegardés puis évalués après une lecture attentive ou un examen scrupuleux.

On utilisait souvent aussi la recherche emboîtée, ou on se servait des liens qu'on trouvait dans les documents examinés pour lancer de nouvelles recherches

Grille d'évaluation d'un site Web[1]

Eléments de l'Evaluation	Critères	Remarques
1. Idenfication du site	l'URL contient (.edu / .org / .gov / .ac.uk)	garantie de la qualité
	Le groupe, l'organisme ou l'établissement qui produit le site	fiable / non fiable
2. Nature et consistance du document	article scientifique / la critique d'un article scientifique	retenu
	article / communiqué de presse	généralement non retenu
3. Longueur du document	court / moyen / long	abandonné / revu / généralement retenu
4. Auteur(s)	identifié? son nom figure-t-il sur le document?	la responsabilité est précisée
	l'auteur: spécialiste du sujet? reconnu dans le domaine? de bonne affiliation?	indice de la qualité
5. Date	document daté? si c'est le cas la date est-elle récente?	c'est une bonne marque
	est-il mis à jour régulièrement?	""

[1] Cette grille a été élaborée en se référant aux documents se trouvant aux adresses suivantes:
http://it2.coe.uga.edu/Faculty/gwilkinson/criteria.html
http://netia59.ac.lille.fr/Ref/pedagogie/Robert_Bibeau/evaluweb.htm
http://www.ccr.jussieu.fr/urfist/cerise/p361.htm
http://csidoc.insa-lyon.fr/sapristi/fristi36.html

6. Référencement	l'existence d'une bibliographie à la fin du document	l'information avancée est certifiée
	le document est cité dans des références connues / se réfère à des références connues	un document important, à retenir
7. Contenu	le document avance du nouveau dans le sujet	c'est une référence importante
	le document constitue une référence pour des documents préalablement retrouvés	indice de la force du document
	document illustré par des tableaux et/ou des graphiques? ces illustrations contribuent-elles à clarifier le texte?	indice de la qualité et de l'effort fourni
8. Structure / couleurs	structure cohérente / titrages utiles / liens pertinents	un bon document
	sobriété des couleurs	un site sérieux

I-2 Les Sites Spécialisées

Vu la multitude de sites spécialisés visités lors de la recherche, et l'impossibilité de les recenser tous, on se suffira à trois de ces sites, tout en détaillant - pour chacun - la méthode employée pour la recherche et l'interrogation.

I-2-1 Findarticles

Est un site spécialisé dans la recherche d'articles en texte intégral enligne, soit dans les revues spécialisées ou celles de presse.

Comme on voit dans la figure dynamique suivante, on peut limiter la recherche à une catégorie de Publications.

Méthode

1) On a limité la recherche à la Catégorie "Computer & Technologie"

2) En suivant les "Search Tips" de FindArticles, on a formulé 3 requêtes de recherche qu'on a utilisées séparément.

- **"Web growth" – commerce – economy – demography – usability – telecommunication**
- **"Indexable Web" OR "Visible Web"**
- **"Invisible Web"**

I-2-2 Librarians Index to the Internet

Portail de ressources Web, élaboré par les bibliothécaires de "Berkely Digital Library SunSITE" sous la direction de "The Library Services and Technology Act" de "California State Librarian" et financé conjointement par "The Library of California" et "U.S. Institute of Museun and Library Services".

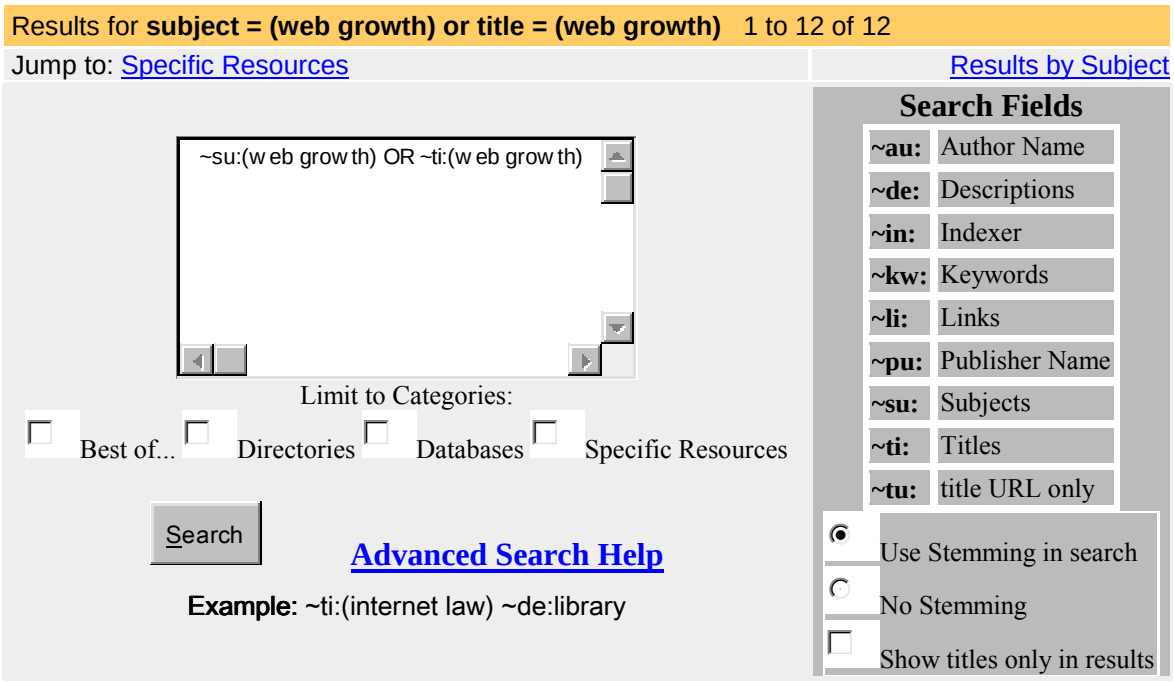
Dans ce Portail, il y a la possibilité:

- de rechercher dans 9 champs différents (dont Auteur ~au:, Titre ~ti:, Liens ~li:, Mots-clés ~kw:, Sujet ~su:, URL ~tu:, ….)
- de limiter la recherche à certaines catégories de ressources (cf. directory, database, etc…)
- d'utiliser les opérateurs booléens

Syntaxe de la recherche, et exemple d'une requête formulée:

~su:(web growth) OR ~ti:(web growth)

= rechercher le termes "web growth" dans les champs du sujet ou du titre



I-2-3 Le Site du NEC Research Institute de Princeton

Un Institut de recherche scientifique, c'est l'établissement d'affiliation de Dr. Steve Lawrence et Dr. Giles, chercheurs dans le domaine de la Recherche de l'Information sur le Web, (Information Retrieval) et auteurs de nombreuses études éminentes sur la taille et la typologie du Web Visible.

Ce site utilise le moteur excite, qui est un outil de recherche paramétrable, largement utilisé dans les sites spécialisés comme moteur intérieur grâce à sa performance et sa capacité de recherche verticale.

Méthode

En employant le moteur intérieur du site on a pu lancer des requêtes très courtes et très précises: **« web growth » « internet growth »**

les résultats ont été affichés sur une page et listés selon le degré (en %) de pertinence à la requête lancée.

La recherche a été affinée par la suite grâce à la possibilité de « cliquer les icônes de certaines références pour trouver des documents similaires » et là le degré de pertinence a

sensiblement augmenté (de 100% pour la même référence jusqu'à 96% pour la moins pertinente) et toutes les références qui en ont résultées ont été retenues.

Cette procédure a été menée avec 2 références jugées les plus pertinentes:

- **Searching the World Wide Web [Steve Lawrence and C. Lee Giles, NEC Research Institute]**
- **Size of the Web, Web Size, Search Engine Coverage and Recency [Steve Lawrence, Lee Giles, NEC Research Institute]**

Les références ont été ensuite sauvegardées et traitées manuellement pour éliminer les doublons.

I-3 Les Bases de Données du Seveur DIALOG

La recherche sur le serveur DIALOG s'est faite sur plusieurs étapes dont certaines ont été préalables à la recherche proprement dite, mais indispensables pour le bon déroulement de celle-ci.

I-3-1 Les commandes et les opérateurs de DIALOG

Ils sont communs et ne rentrent pas dans les particularités de chaque base.

Sans prétendre à l'exhaustivité, le tableau suivant résume les commandes et les opérateurs qu'on a utilisés au cours de la recherche.

Commandes		
	begin (b)	• ouvrir la/les base(s) dont les noms ou les No. suivent • ouvrir les bases appartenant à une catégorie
	select (s)	• cherche le/les termes qui suivent • combiner plusieurs étapes de recherche
	remove duplicate (rd)	éliminer les doublons si présents
	type (t)	visualiser les résultats d'une recherche
	sort	trier les notices dans l'ordre croissant d'un ou de plusieurs champs
Opérateurs booléens		
	AND / OR / NOT	
Opérateurs de proximité		
	W	rechercher des mots adjacents dans un ordre strict
	N	rechercher des termes voisins dans un ordre indifférent
Troncature		
	?	remplace un caractère à l'intérieur d'un mot / un ou plusieurs caractères à la fin

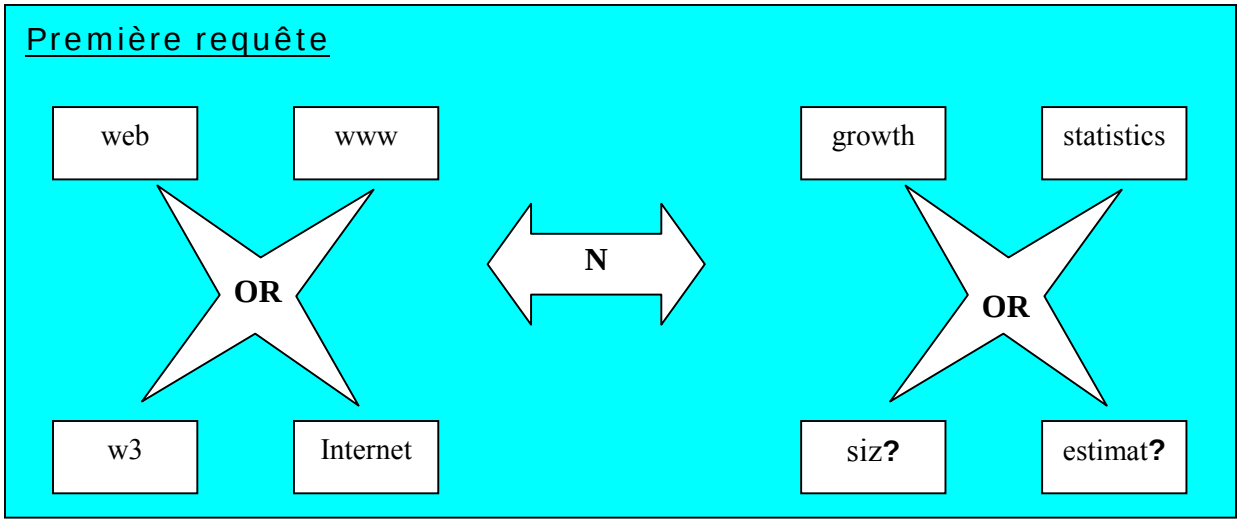
I-3-2 La Formulation des Requêtes et l'Utilisation de chaque Opérateur

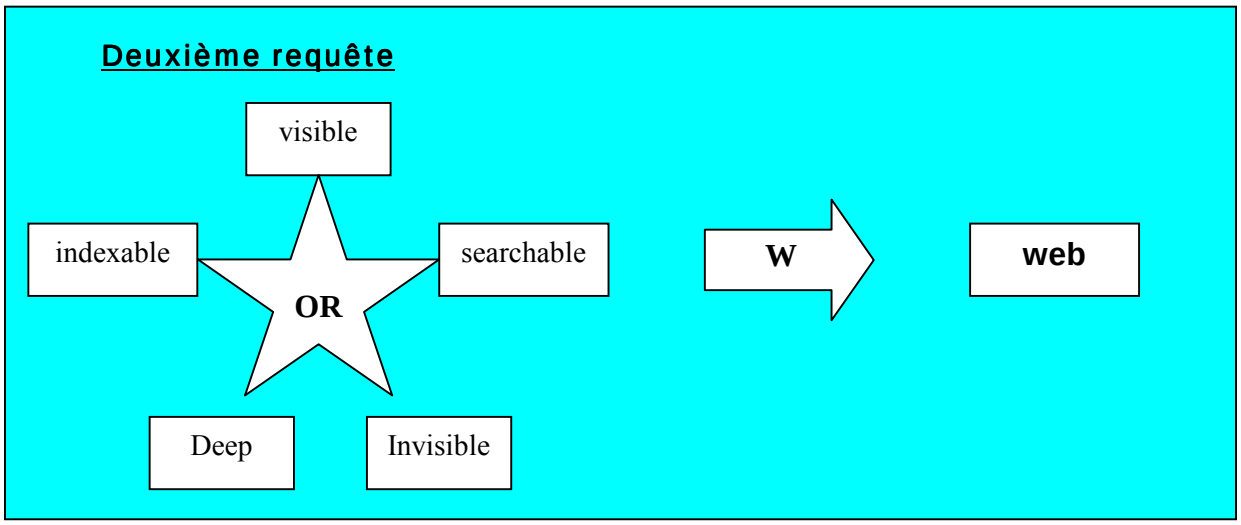
Dans la formulation des requêtes ont été pris en considération:

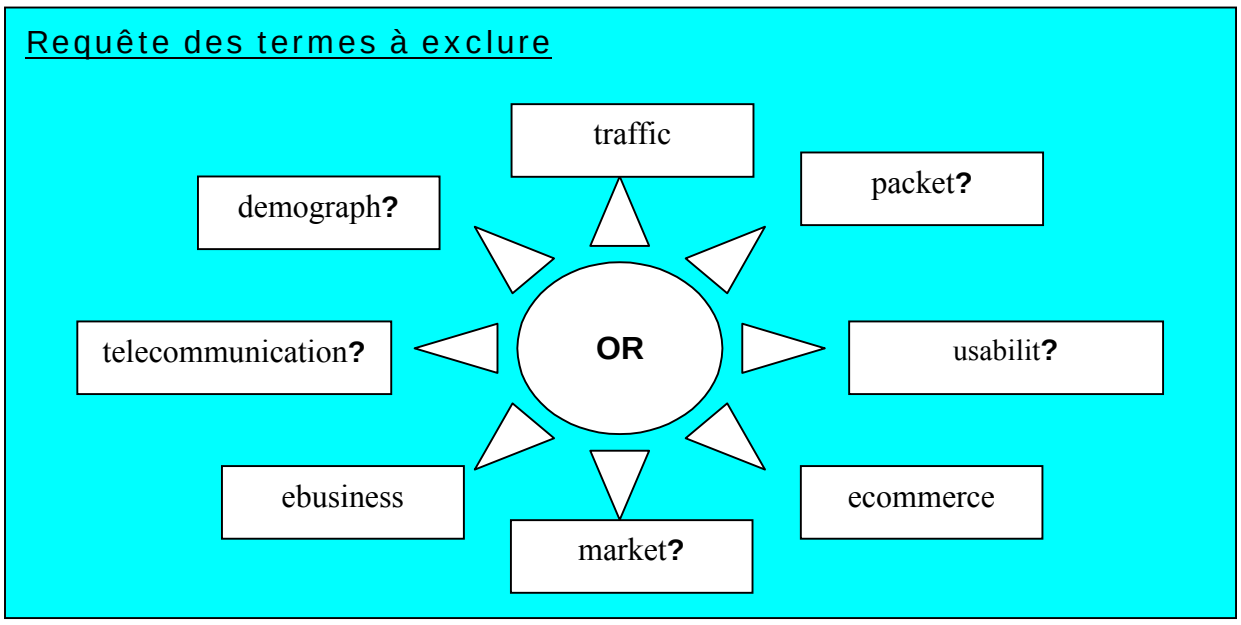
- les différentes alternatives des mots importants ainsi que les différentes variantes sémantiques **"OR"**
- les mots de la même famille **"?"**
- l'ordre des termes de la requête selon qu'il est strict **"W"** ou indifférent **"N"**

Trois requêtes ont été formulées dont une pour les termes à exclure[1]

[1] En réalité les termes à exclure sont plus nombreux que ceux figurant ici à titre explicatif


Première requête
web
www
OR
w3
Internet
N
growth
statistics
OR
siz?
estimat?


Deuxième requête
visible
indexable
searchable
OR
Deep
Invisible
W
web


Requête des termes à exclure
traffic
demograph?
packet?
telecommunication?
OR
usabilit?
ebusiness
ecommerce
market?

I-3-3 Sélection des bases de données à interroger

1.1.1.2. Les One Search Categories

Pour sélectionner la majorité des bases à interroger on a utilisé la fonction **"One Search Catégories"** de DIALOG qui permet de rechercher simultanément plusieurs bases qui ont été groupées sous une catégorie qui est ici: **"Science - Computer Technology"**.

Trois **"Search Groups"** font partie de cette catégorie: [COMPSCI] [SOFTWARE] [PCINFO]. Cependant quelques modifications ont été effectuées comme suit:

- le 3e groupe a été abandonné, car certaines de ses bases sont déjà présentes dans le 1er groupe, d'autres ne sont pas pertinentes pour notre sujet.
- par contre la base LISA (Library and Information Science Abstracts [61]) en relation très étroite avec le thème et n'étant pas présente dans aucun des groupes sélectionnés, a été ajoutée.
- les bases non pertinentes, comme [92]; [108]; [238]; [434] du 1er groupe et [256] du second, ont été exclues
- la base No. [34] du 1er groupe également a été exclue, quoique pertinente, pour des raisons que nous allons élucider plus loin
- restent deux bases, la [674] et la [103] que le serveur a refusé d'ouvrir lors de la recherche

En ce qui suit, les bases retenues dans chaque groupe:

DIALOG
A THOMSON COMPANY

Selected Databases

Databases by Subject Category:
Science - Computer Technology

COMPSCI | SOFTWARE

Computers, Electronics, and Telecommunications [COMPSCI]

- CMP Computer Fulltext [647]
- Conference Papers Index [77]
- DIALOG Telecommunications Newsletters [696]
- Dissertation Abstracts Online [35]
- Ei Compendex® [8]
- Energy Science and Technology [103]
- Gale Group Computer Database™ [275]
- INSPEC (1969-present) [2]
- Information Science Abstracts [202]
- Inside Conferences [65]

- Internet & Personal Computing Abstracts™ [233]
- JICST-EPlus - Japanese Science & Technology [94]
- MathSci® [239]
- NTIS - National Technical Information Service [6]
- PASCAL [144]
- Wilson Applied Science & Technology Abstracts [99]

Software Directories [SOFTWARE]

- Internet & Personal Computing Abstracts™ [233]
- MicroComputer Software Guide OnLine™ [278]

Les fiches techniques (Blue Sheets)

Un temps à été consacré à la consultation des fiches techniques des bases retenues (une fiche par base) pour connaître les particularités de chacune d'elles dont:

- les sujets et la période couverts par chacune
- les indexes implicites de chaque base
- les indexes qui peuvent être utilisés comme suffixes servant à limiter la recherche à certains champs

1.1.1.3. Les Bases Recherchées Conjointement

- Social SciSearch® - 1972- [7]
- SciSearch® - a Cited Reference Science Database - 1990- [34][1]

Ces deux bases ont été interrogées conjointement pour les raisons suivantes:

- elles ont à peu près les mêmes particularités
- leurs indexes implicites sont presque les mêmes
- les deux bases étant multidisciplinaires, elles offrent une possibilité d'affinage (SC = find a subject category) qui nous a permis de limiter la recherche aux seules catégories ayant trait à notre sujet, ce qui a réduit considérablement le nombre de notices récupérées.[2]

[1] C'est la base qui a été exclue du 1er groupe

[2] 9 notices, alors que sans cet affinage on récupérait 49 notices supplémentaires pour qui l'expression "Web Size" signifie le diamètre des toiles d'araignées

I-3-4 Interrogation des Bases

I	Les One Search Categories	Remarques
	B COMPSCI, SOFTWARE, 61, NOT 34, 92, 108, 238, 434	
Ouvrir les bases contenues dans COMPSCI, SOFTWARE en ajoutant la base [61] et en excluant les bases No. [34], [92], [108] [434], [238]		
S1	**S** (VISIBLE OR INDEXABLE OR SEARCHABLE OR INVISIBLE OR DEEP) (W) WEB/TI, AB, DE, ID, NT, SH	
S2	**S** (INTERNET OR WEB OR WWW OR W3) (N) (STATISTICS OR GROWTH OR SIZ? OR ESTIMAT?)/TI, AB, DE, ID, NT, SH	
Rechercher les termes qu'il faut retenir tout en limitant la recherche aux champs du (TI) Titre, (AB) Résumé, (DE) Descripteurs, (ID) Mots-clés d'auteurs, (NT) et (SH) Subject Headings (ces trois derniers champs n'existent pas dans toutes les bases recherchées)		
S3	**S** TRAFFIC OR TELECOMMUNICATION? OR PACKET? OR CONNECTIVITY OR DEMOGRAPH? OR USAGE? OR USABILIT? OR JAPAN OR CHINA OR ARGENTINA OR (ECONOMIC (W) ASPECT?) OR EBUSINESS OR ECOMMERCE OR MARKET? OR (ELECTRONIC (W)(BUSINESS OR COMMERCE))	***Recherche faite dans le but d'un affinage***
Rechercher les termes qu'il faut exclure quel que soit leur emplacement dans les notices		
S4	**S** (S1 OR S2) NOT S3	***1er affinage***
Unir les notices résultantes des recherches S1 et S2 et exclure celles résultantes de la recherche S3		
S5	**S** S6/1998:2001	***2ème affinage***
Appliquer le critère de la date pour diminuer le nombre de notices		
S6	**RD** S5	***3ème affinage***
Supprimer les doublons qui peuvent exister dans les bases pour ne retenir que des références uniques		
	T S6/9/all	
Afficher les notices dans un format complet et les sauvegarder pour les examiner et juger de leur pertinence		
S7	**SORT** S6/AU, JN, TI	***Pour faciliter l'organisation des références dans la bibliographie***
Trier les notices selon (1) le nom de l'Auteur ensuite (2) le titre du périodique et enfin (3) le Titre de l'Article		
	T S7/AU, TI, JN, PY, SO	
Afficher les notices dans un format personnalisé avec les champs: (1) Auteur, (2) titre de l'article, (3) titre du périodique, (4) date de publication		

II	Les Bases Interrogées conjointement	Remarques
	B 7, 34	
Ouvrir les bases [7] et [34]		
S1	**S** SC=COMPUTER APPLICATIONS OR CYBERNETICS OR (INFORMATION OR LIBRARY (W) SCIENCE) OR ELECTRONICS	
Limiter la recherche à ces catégories pour ne pas inclure les autres disciplines qui peuvent générer du bruit		
S2	**S** (INTERNET OR WEB OR WWW OR W3) (N) (STATISTICS OR GROWTH OR SIZ? OR ESTIMAT?)/TI, AB, DE, ID	
Rechercher les termes qu'il faut retenir tout en limitant la recherche aux champs du (TI) Titre, (AB) Résumé, (DE) Descripteurs et (ID) Mots-clés d'auteurs		
S3	**S** S1 AND S2	***1er affinage***
Combiner les catégories choisies (en S1) avec la requête de (S2) pour ne pas avoir du bruit ou des notices non pertinentes provenant d'autres catégories pour qui le terme Web signifie araignée (l'insecte)		
S4	**S** (VISIBLE OR INDEXABLE OR SEARCHABLE OR INVISIBLE OR DEEP) (W) WEB/TI, AB, DE, ID	
Même procédure qu'en S2		
S5	**S** TRAFFIC OR TELECOMMUNICATION? OR PACKET? OR CONNECTIVITY OR DEMOGRAPH? OR USAGE? OR USABILIT? OR JAPAN OR CHINA OR ARGENTINA OR (ECONOMIC (W) ASPECT?) OR EBUSINESS OR ECOMMERCE OR MARKET? OR (ELECTRONIC (W)(BUSINESS OR COMMERCE))	***Recherche faite dans le but d'un affinage***
Rechercher les termes qu'il faut exclure quel que soit leur emplacement dans les notices		
S6	**S** (S3 OR S5) NOT S4	***2ème affinage***
Unir les notices résultantes des recherches S3 et S4 et exclure celles résultantes de la recherche S5		
S7	**S** S6/1998:2001	***3ème affinage***
Appliquer le critère de la date pour diminuer le nombre de notices		
S8	**RD** S7	***4ème affinage***
Supprimer les doublons qui peuvent exister dans les deux bases		
	T S9/9/all	
Afficher les notices dans un format complet et les sauvegarder pour les examiner et juger de leur pertinence		
S9	**SORT** S9/AU, JN, TI	***Pour faciliter l'organisation des références dans la bibliographie***
Trier les notices selon (1) le nom de l'Auteur ensuite (2) le titre du périodique et enfin (3) le Titre de l'Article		
	T S11/AU, TI, JN, PY, SO	
Afficher les notices dans un format personnalisé avec les champs: (1) Auteur, (2) titre de l'article, (3) titre du périodique, (4) date de publication		

I-3-5 Analyse des Résultats Obtenus

Nombre de notices obtenues	96
Le Web Visible et les moteurs de recherche	**15**
Typologie du Web	**1**
Analyse de la croissance du Web	**6**
Web Invisible et outils de recherche	**12**
Les sites miroirs et leur nombre	**1**
Total des notices retenues	35
Nombre des notices non-pertinentes	61

Provenance:
"Internet growth"
"Web statistics" "Internet Statistics"

I-3-6 Coûts d'Interrogation

	FF
Les One Search Categories	129
Les bases [7] et [34]	29.6
TOTAL	**158.6**

I-4 Les Journaux Electroniques au Site de l'ENSSIB

Les signets vers quelques journaux spécialisés enligne accessibles depuis le site de l'ENSSIB, ont constitué une source très appréciée de références. Dans ce qui suit, on présentera quelques exemples de ces journaux et on expliquera comment on a pu obtenir des résultats.

I-4-1 CyberMetrics

International Journal of Scientometrics, Informetrics and Bibliometrics
ISSN 1137-5019

Revue électronique spécialisée en scientometrie, Informétrie et bibliométrie

A partir de l'**INDEX** du Journal, (situé toujours dans la colonne de gauche constituant l'axe de navigation du site), la sous-rubrique **TOOLS** de la rubrique **The Source** a été notre principale source d'information grâce à ses deux premiers liens: "Measuring the Net", et "Searching the Web"

The Source >TOOLS>Measuring the Net
The Source >TOOLS>Searching the Web

Ces liens pointent vers des pages compilant des signets de sites extérieurs, jalonnés de quelques commentaires et d'un nombre de synthèses par les rédacteurs du journal.

I-4-2 FreePint

Free Pint ISSN : 1460-7239

Revue se concentrant sur les techniques de recherche et les ressources sur Internet.

En utilisant le **Topic Index**, on a eu accès au Portail du site, dans lequel les articles sont classés suivant différentes catégories.

En choisissant la catégorie "Information and Libraries", on a pu localiser quelques articles.

I-5 La Recherche Manuelle

Un dépouillement des articles des périodiques en papier de la bibliothèque de l'ENSSIB a été effectué.

Cette recherche vient en 5e lieu après:

- la recherche sur le web via les moteurs de recherche
- la recherche des sites spécialisés sur le web
- l'interrogation des BDD de DIALOG
- les journaux électroniques

I-5-1 Les Critères de la Recherche

C'est la recherche la plus longue et la plus minutieuse; la plus longue dans le sens qu'elle nécessite beaucoup de temps, et plus minutieuse dans le sens que le traitement se fait en même temps que la recherche.

Trois critères nous guidaient dans cette recherche:

- sélectionner les revues spécialisées en Internet et en sciences de l'Information
- exclure parmi ces dernières, les revues qui ont été recherchées sur le Web ou le serveur DIALOG (cf. LISA, ONLINE)
- dépouiller tous les numéros des revues sélectionnées à partir de début 1999 jusqu'aux derniers numéros disponibles à la Bibliothèque de l'ENSSIB (généralement début 2001, ou fin 2000 pour certaines revues).

Méthode

- Consulter les sommaires des revues sélectionnées à la recherche d'un article ou d'une rubrique se rattachant à notre sujet
- les articles "soupçonnés" sont examinés pour juger de leur pertinence et de leur consistance; plusieurs éléments sont pris en considération dont: la longueur de l'article, sa structure, le titrage des différentes parties, les illustrations, la bibliographie et les références.
- Consulter les rubriques "Calendriers" ou "Coming Events" et suivre les liens qu'ils donnent sur le Web, pour retrouver des conférences ou des colloques.

I-6 Le Silence Français

Plusieurs recherches ont été lancées à plusieurs reprises sur Voilà, Nomade et Google.fr en utilisant des mots-clés français:

"taille du Web"	**"croissance du Web"**
"Web Visible"	**"Web Invisible"**

les résultats sont très décevants.

- Pour les deux premières requêtes une seule référence, qui n'a pas été retenue à cause de sa faiblesse, sa superficialité et sa mal interprétation de l'étude du bureau de recherche de l'OCLC pour estimer la taille du Web Visible.
- Pour les deux dernières requêtes un tas de références dénombrant des outils, des sites et des "astuces" (pour conserver leur terme) pour rechercher le Web Visible et surtout le Web Invisible, en plus d'un grand nombre de sites proposant des formations – payantes bien sûr - à l'utilisation de ces outils et "astuces".

I-7 Critères Généraux de Sélection des Références

La deuxième étape de la recherche sur le Web nous a permis d'établir des critères généraux pour la sélection des références retrouvées:

Ont été retenues:

- Les études qui estiment la taille et la croissance du Web en nombre de pages ou de sites
- Les études qui s'intéressent au dynamisme du Web ou à sa typologie
- Les articles qui ont expliqué ou commenté ou critiqué ces études
- Les références qui s'intéressent à la couverture du Web par les moteurs de recherche
- Les documents qui traitent des problèmes que posent la taille et la croissance du Web à l'accès à l'information

- Les références qui traitent de la recherche et l'extraction de l'Information sur le Web (IR = Information Retrieval or IE = Information Extraction) en général sans viser certains outils
- Les références qui parlent de la typologie du Web Visible ou du Web Invisible ou qui en donnent une définition ou en estiment l'étendue
- Les références qui recensent les domaines et les hôtes sur le Web

Ont été abandonnées:

- Les références qui comparent la taille des différents moteurs de recherche sans évoquer leur couverture du Web.
- Les documents consacrés à la description d'outils spécifiques de la Recherche d'Information sur le Web (IR) (cf. BullsEye d'IntelliSeek, etc,…)
- Les documents trop courts ou les articles de presse qui rapportent succinctement les chiffres ou les résultats de certaines études déjà trouvées.
- A l'exception de certaines études vraiment remarquables, aucune référence datant d'avant 1998 n'a été retenue.

I-8 L'Accès aux Documents Primaires.

Les documents primaires qui ont servi à la rédaction de la synthèse, ont été obtenus par divers moyens, on en distingue plusieurs catégories:

1. Les documents en HTML qui ont été obtenus lors de la recherche sur le Web via les différents moteurs de recherche.
2. Les articles de périodiques qui ont été photocopiés lors de la recherche manuelle.
3. Quant aux documents primaires correspondant aux références résultant de la recherche des bases de données du serveur DIALOG, on a pu obtenir leur version électronique grâce à l'une des trois manières suivantes:

1.1.1.4. sur Google

en lançant une recherche sur le moteur Google, composées du titre de l'article écrit entre guillemets suivi de l'indication PDF, car ce moteur est capable de rechercher aussi les documents en ce format

1.1.1.5. sur Findarticles.com

en choisissant le titre de la revue ou du journal désiré, puis en tapant le titre de l'article – toujours entre guillemets – dans le formulaire de recherche.

1.1.1.6. sur les sites des revues enligne

pour les revues qui permettent un accès au texte intégral de leurs numéros, on lançait une recherche par titre d'article ou par auteur sur le moteur intérieur du site, ou on allait directement sur le numéro qui contient l'article recherché.

I- 9 Temps Consacré à l'Ensemble de la Recherche

	Heures
Initiation aux différents outils (fiches techniques, langage et méthodes d'interrogation, informations diverses) + formulation des requêtes pour chaque outil	**12**
Recherche sur le Web	**60**
Recherche sur Dialog	**1**
Recherche Manuelle	**11**
Traitement des références obtenues et lecture	**28**
Référencement et élaboration de la Bibliographie et vérification des liens	**20**
Rédaction	**24**
TOTAL	**136**

I- 10 Remarques générales sur la recherche

Avant de quitter cette partie, il nous paraît important de dégager quelques remarques sur la recherche dans son ensemble, comme sur les difficultés rencontrées lors de cette recherche et qui se sont répercutées sur les diverses parties du travail.

Ces difficultés sont de plusieurs ordres et on peut les résumer ainsi:

1. Il y a des périodes où la recherche sur le Web stagne et commence à tourner en rond, ce qui fait que même en multipliant les requêtes et en diversifiant les méthodes d'interrogation on ne tombait pas sur de nouveaux documents (problème dû aux délais importants de mise à jour des indexes des moteurs).

2. Ceci a eu comme conséquences: un temps d'attente important, et une insatisfaction par rapport aux résultats obtenus, car ceux-ci ne pouvaient pas constituer le noyau d'une bibliographie, ni la base d'une rédaction (même le plan ne pouvait pas être imaginé!)

3. Le temps consacré à la recherche manuelle n'a pas été bien calculé; comme il a été décidé de terminer par cette recherche – ce qui semble la démarche la plus logique- on l'a faite à la fin juste pour compléter la bibliographie, sans croire qu'elle pourrait ouvrir de nouveaux horizons pour la recherche sur le Web par le biais des rubriques d'évènements (cf. conférences, colloques) qui constituent une source inestimable de références, ce qui a fait qu'au lieu de clore la recherche, on l'a relancée et, dans un temps critique!

4. On a été confronté à des difficultés non négligeables pour le référencement de sites ou de pages Web, surtout que, dans certains cas, l'auteur ou/et la date du document n'y figuraient pas, ainsi que parfois le titre n'est pas très évident, sans compter aussi le temps passé à rechercher ces informations.

5. L'absence de références françaises sur le sujet, a multiplié le temps de traitement et d'analyse de celles-ci, ainsi que le temps de consultation des documents primaires, qui dans la plupart du temps abordent la question d'un point de vue très scientifique et avec des détails techniques et en anglais!!!

II- SYNTHESE

II- SYNTHESE

La taille du Web, sa croissance, sa typologie, ont fait couler beaucoup d'encre: depuis les articles scientifiques jusqu'aux articles de presse, passant par les critiques et les "peer reviews". Cependant le nombre d'études-phares traitant du sujet est relativement bas. On vise par études phares, les études qui avancent de nouveaux chiffres, qui présentent à elles seules une nouveauté et dont la méthodologie employée a été bien définie et expliquée. Ces études sont au nombre de cinq, et on essayera dans ce qui suit de les analyser et de présenter les résultats sur les lesquels elles ont conclu.

Malgré l'interface uniforme qu'il affiche et la navigation aisée et continue que l'utilisateur sent allant de lien en lien, le Web n'est pas une seule entité caractérisée par l'homogénéité et la cohérence. Il existe deux entités distinctes du Web impliquant chacune un certain nombre de paramètres et faisant appel à des technologies différentes de recherche et de traitement de l'information. **[56]** Cette dichotomie est communément exprimée par deux termes qu'on rencontre souvent: le "Web Visible" et le "Web Invisible".

II-1 Le Web Visible ou "Indexable par les Moteurs"

- La définition la plus simple qu'on peut donner du "Web Visible" est l'ensemble des fichiers HTML hébergés dans les serveurs Web, et qui peuvent être visionnées au moyen d'un navigateur (le Netcape ou l'Explorer pour ne citer que les plus connus). Pour plus de simplicité le terme "fichiers HTML" a été détrôné par le terme "Pages Web" plus utilisé et largement diffusé.
- Ces pages sont constituées essentiellement d'un texte[1] et de balises HTML servant, d'une part, à mettre en page ce texte, d'une autre part, à lier ces pages entre elles.
- Le terme "Visible" illustre leur état par rapport aux moteurs de recherche. En fait, pour qu'une page soit "visible" à un robot, elle doit être "statique" et "navigable". "Statique" dans le sens que sa présence sur le Web ne dépend d'aucun autre facteur que la volonté du Webmaster qui l'a créée et rendue publique: elle est toujours là, affichant le même contenu à tout le monde, sauf si son créateur a décidé de la déplacer ou de changer son contenu. "Navigable" dans le sens qu'elle est liée à d'autres pages à travers des liens hypertextes pointant vers elle ou sortant d'elle.
- Rentrent dans cette catégorie:
  - les pages constituées d'hypertexte et d'hyper liens
  - les pages construites sur le principe de l'arborescence ou la catégorisation, (car qu'est ce que l'arborescence sinon un nombre de liens)
  - les pages "stop", c'est à dire celles qui ne comportent aucun lien vers l'extérieur ou qui ont un seul lien vers la page d'accueil de leur site

[1] Dans la définition que l'OCLC donne de ce qu'il appelle "la plus petite unité du Web", qui est la page Web, le texte est la composante essentielle de la page, tout autre élément comme l'image, le son, etc… est considéré comme supplémentaire. **[60]** (en réalité il ne faut pas confondre le HTML et le numérique).

- et le cercle s'élargit pour englober toutes les pages accessibles librement sur le Web sans restriction ni exigence d'une identification préalable. **[61] [62]**

On étalera dans ce qui suit, - dans l'ordre chronologique bien entendu -, quatre études qui se sont intéressées à estimer la taille du Web Visible en nombre de pages, ou de sites.

II-1-1 L'Estimation de la Taille du Web par le Chevauchement entre les Moteurs de Recherche - décembre 1997 [62]

L'objectif principal de cette étude réalisée par Dr. Steve Lawrence et Dr. Lee Giles, les chercheurs au NEC Research Institute de Princeton, était d'évaluer la performance des moteurs de recherche et de comparer la taille de leurs indexes respectifs en vue d'estimer leur couverture du Web Invisible. Ensuite le chevauchement entre chaque paire des moteurs testés a été analysé pour estimer la taille du Web. **[62]**

Méthodologie

- 575 requêtes collectées auprès du personnel du NEC Research Institute et élaborées par nos deux chercheurs, ont été posées aux 6 grands moteurs de l'époque[1] durant la période du 15 au 17 décembre 1997.
- Les résultats ont été traités et analysés et la couverture de chaque moteur a été comparée à la couverture combinée des six moteurs, en vue de les classer du plus grand au plus petit.
- Ensuite, le chevauchement entre les paires de moteurs (des 2 les plus petits aux 2 les plus grands) a été calculé pour estimer à chaque fois la fraction du Web susceptible d'être indexée **(cf. Table 1)**.
- Cette méthode est rendue peu fiable par la "dépendance" des moteurs entre eux, c'est à dire la duplicité et l'existence de doublons, car le contenu de chaque moteur n'est pas unique: du fait que les utilisateurs enregistrent leurs pages auprès de plusieurs moteurs; du fait aussi que les moteurs eux-mêmes sont enclins à indexer les pages les plus populaires et les plus recherchées par les utilisateurs.
- Cette constatation a été déterminante dans le choix des deux moteurs dont le chevauchement estimera la taille du Web. On s'attend logiquement à ce que les moteurs aux plus larges indexes aient une faible dépendance. En fait ces moteurs sont ceux qui ont affiché les meilleurs résultats en réponse aux requêtes scientifiques formulées par le personnel de l'Institut, l'information scientifique étant connue pour être la moins populaire et la plus difficile à trouver.

[1] Les 6 moteurs dans l'ordre alphabétique sont:
AltaVista, Excite, HotBot, Infoseek, Lycos, et Northern Light

Les paires de moteurs	Web Indexable (par millions de pages)
Lycos et Infoseek	90
Infoseek et Excite	220
Excite et Northern Light	230
Northern Light et AltaVista	230
AltaVista et HotBot	***320***

Table 1: Le chevauchement par paires de moteurs – tableau traduit et reproduit d'après [108] [62]

Résultats de l'étude

- La taille du Web a été estimée en décembre 1997 à 320 millions de pages
- le nombre de pages uniques résultants de la combinaison des résultats des 6 moteurs, après avoir éliminé les doublons, est de 190 millions de pages.

II-1-2 L'Echantillonnage du Web et le Test d'Adresses IP

La taille exacte du Web est inconnue, le Web est certes trop vaste pour permettre une analyse exhaustive de son contenu. L'approche la plus pratique et la plus raisonnable est de collecter un échantillon représentatif du contenu du Web, de le tester et ensuite extrapoler les résultats obtenus à l'ensemble[1]. La fiabilité et l'exactitude des estimations reposent sur le processus et la qualité de l'échantillonnage. En particulier, l'échantillon doit être choisi aléatoirement pour que chaque élément ait la chance d'être sélectionné.

Il est nécessaire de développer une méthodologie pour la collecte de l'échantillon à travers d'adresses IP produits aléatoirement. L'adresse IP est l'équivalent numérique du "top level" URL ou l'URL de base correspondant à la page d'accueil d'un site. Cette méthode vise donc à localiser aléatoirement des sites Web, dont les caractéristiques seront étudiées par la suite. C'est un moyen indirect de collecter des pages Web, puisque chaque site est constitué d'un ensemble de pages, donc si un échantillon de sites Web est obtenu, par défaut un échantillon de pages Web est aussi obtenu**.[105]**.

[1] Dans un article intitulé *Growth Dynamics of the World-Wide Web*, les auteurs attirent notre attention sur le fait que "dans cette écologie de savoir, qu'est le Web, dans laquelle diverses sortes d'information sont liées de manière extrêmement complexe et arbitraire, il existe quand même un ordre latent gérant l'ensemble: les pages sont distribuées à travers les sites suivant une "loi universelle de puissance": plusieurs sites ont peu de pages, tandis qu'une petite minorité de sites ont des centaines de milliers de pages. Cette loi gérant la croissance du Web, rend possible la détermination de sa taille en se basant sur n'importe quel nombre de sites sans avoir à parcourir tout l'ensemble". **[58]**

1.1.1.7. L'Etude de Dr. Lawrence et Dr. Giles – février 1999 [61] [108]

- Durant la période allant du 2 au 28 février 1999, les auteurs ont choisi des adresses IP aléatoires et testé la présence d'un serveur Web au Port standard (le 80 est le Port standard par défaut du protocole HTTP).
- Parmi un total de 4.3 milliards d'adresses IP dans le cyberespace, une fraction de 3.6 millions a été testée à la recherche de serveurs; 1 serveur tous les 269 adresses a été trouvé, soit 16 millions de serveurs Web en tout.
- Cette estimation n'a pas pour autant constitué une base solide pour le calcul du volume d'information sur le Web Visible, car ce nombre inclut des serveurs ne faisant pas partie du Web Indexable par les moteurs, tels les serveurs exigeant des autorisations (y compris des murs de feu), des serveurs répondant par des pages standard et ceux sans contenu (les sites "coming soon"), les imprimantes, les routeurs, les proxies, les serveurs de messagerie et de CD-ROM et tous les parcs informatiques se dérobant derrière une interface Web.
- Une BDD pour les expressions permettant d'identifier ces serveurs a été construite. Ensuite il a été procédé à une classification manuelle de tous les serveurs, tout en supprimant ceux qui ne font pas partie du Web Visible.
- Les sites ayant le même contenu sur multiples adresses IP ont été comptés pour un seul chacun.
- Les résultats obtenus estimaient le nombre de serveurs du Web dit indexable à 2.8 millions, sachant qu'un serveur donné peut héberger plusieurs sites.
- Pour estimer le nombre total de pages du Web Indexable, les auteurs ont parcouru toutes les pages des premiers 2500 serveurs Web choisis aléatoirement.

Résultats de l'étude

Les tableaux suivants résument les résultats de l'étude

Nombre moyen de pages par serveur	**289**
Nombre moyen d'images par serveur	**62,8**

	Médiane	Moyenne
Taille de la page (Ko)	**18,7**	**3,9**
Contenu textuel (Ko)	**7,3**	**0,93**
Taille de l'image (Ko)	**15,2**	**5,5**

Volume Total de:	Pages	Contenu textuel	Données iconographiques
En tera octets	15	6	3

Nombre total d'images	180.000.000

Nombre total de pages Web	800.000.000

La Figure 1 illustre la distribution des sites du Web Visible et le pourcentage qu'occupe chaque domaine.

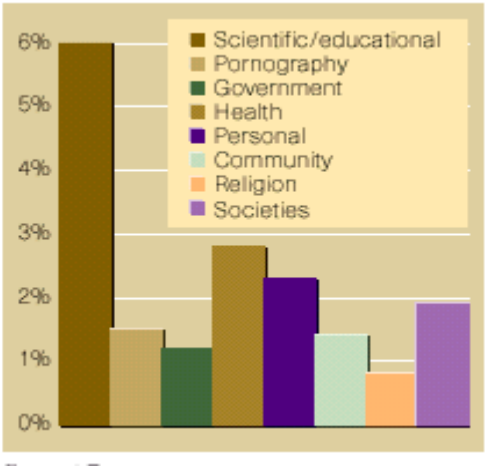


Figure 1 Distribution de l'information sur le Web Visible en février 1999. Environ 83% des sites sont commerciaux. Le reste est illustré dans la figure.
Figure copiée de la version PDF de l'article "Accessibility of the Information on the Web". [61]

L'étude a été critiquée pour le chiffre "très arrondi" qu'elle présente (800 millions de pages); [42] en fait, les auteurs avouent que ce nombre peut être faussé et que la valeur réelle de la taille du Web peut être plus élevée grâce aux sites rarissimes qui sont constitués de millions de pages (cf. Geocities 34 millions de pages), ou parce que certains sites n'ont pas été entièrement parcourus au moment de l'étude à cause d'erreurs survenues. **[61]**

1.1.1.8. L'Etude de l'OCLC – 2000 [105]

Si l'étude de Dr. Lawrence et Dr. Giles ne nous renseigne pas sur la méthode d'obtention des adresses IP testées[1], par contre celle de l'OCLC nous détaille la méthodologie employée pour ce faire.

- L'adresse IP est une séquence de 32 bits qui identifient une connexion à Internet. Elle est toujours exprimée sous une forme décimale: 4 octets de 8 bits chacun, séparées par des points; chaque octet est converti en un nombre décimal comme suit:

10000100 . 10101110 . 00000001 . 00000101

132 . 174 . 1 . 5

- Comme chaque octet est composé de 8 bits, la conversion décimale peut prendre une valeur allant de 0 jusqu'à 255 inclus. Quant au nombre du Port, c'est un identificateur de 16 bits qui ajoute un degré de spécificité à une adresse Internet, car une connexion identifiée par une adresse IP peut comprendre simultanément plusieurs types de service Internet, il est donc nécessaire de préciser auquel de ces services une transmission de données est dirigée: dans notre cas pour le protocole HTTP c'est le Port 80.
- Le cyberespace contient actuellement 4.294.967.296 adresses IP[2], dont chacun est susceptible d'héberger un ou plusieurs services Internet (mél, FTP, Goopher, etc…) En plus, une bonne partie de ces adresses IP n'est pas assignée; on en déduit que les adresses IP correspondant à des sites Web représentent une fraction du nombre total d'adresses.
- En vue de minimiser la perte du temps et de ressources, il est utile de restreindre ce vaste espace d'adresses IP en excluant les tranches d'adresses connues pour êtres invalides ou non utilisées. Il existe pour ceci des informations autoritaires concernant les séries d'adresses que l' "Internet Assigned Numbers Authority" (IANA) n'a pas attribuées aux utilisateurs. Sachant que ces tranches d'adresses ne contiennent pas d'hôtes accessibles, ont peut les exclure du champ de l'expérience sans problème.[3] Ces éliminations constituent environ 48% du total des adresses IP, ils en restent donc 2.230.124.544.Une fois les dimensions de notre champ d'étude délimitées, il ne reste qu'à produire aléatoirement des adresses IP.
- Grâce à leur forme numérique, ces adresses peuvent êtres produites aléatoirement au moyen d'un RNG (Random Number Generator = générateur de nombres aléatoires) paramétré pour exclure les adresses non assignées.

---

[1] "We have now obtained and analyzed a random sample of servers […]. There are currently 254^4 possible IP addresses […] some of these are unavailable, while some are known to be unassigned."**[61]**

[2] Avec la IPv6, la nouvelle version du protocole IP ce nombre augmentera considérablement. **[61]**

[3] Ces tranches correspondent à des hôtes non accessibles, des portions d'adresses IP qui n'ont pas été attribuées par IANA, et celles réservées à des réseaux privés sans connexion externe vers l'Internet. **[105]**

Résultats de l'étude:

Le bureau de recherche de l'OCLC a pris le soin de suivre la croissance du Web depuis 1997 jusqu'à 2000. Les tableaux et les graphiques suivants ont été conçus d'après les chiffres de l'étude. **[107]**

Taux de croissance

	'97 - '00	'97 - '98	'98 - '99	'99 - '00
Sites Web	371%	82%	71%	52%
Sites Uniques	---	---	77%	53%
Sites publics uniques	268%	82%	53%	32%

Volatilité d'adresses

Pourcentage d'adresses IP qui ont identifié des sites en une année et ce qui en persiste l'année suivante

	1998	1999	2000
1998	**100%**	**56%**	**35%**
1999	**---**	**100%**	**55%**
2000	**---**	**---**	**100%**

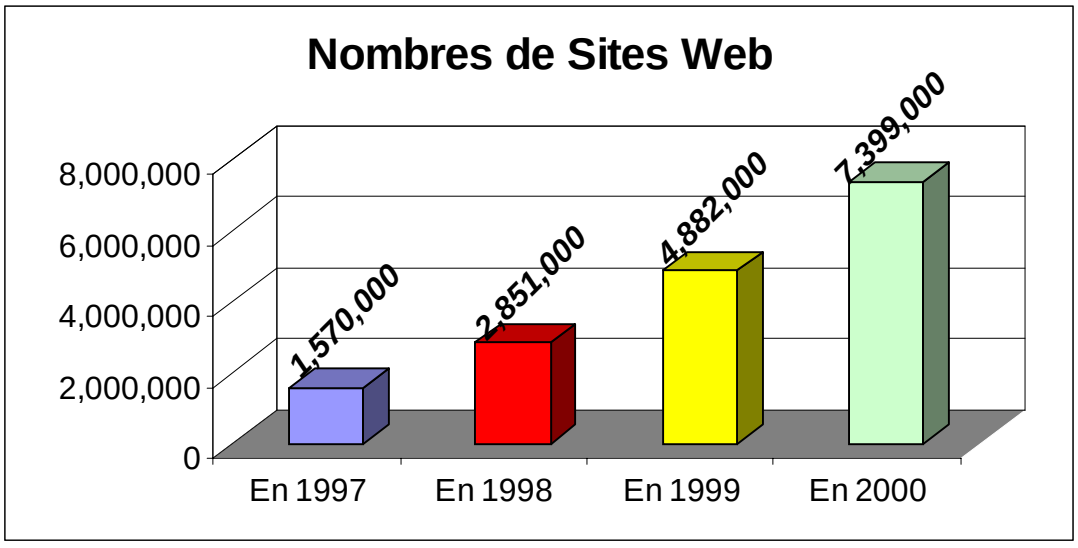


Figure 2

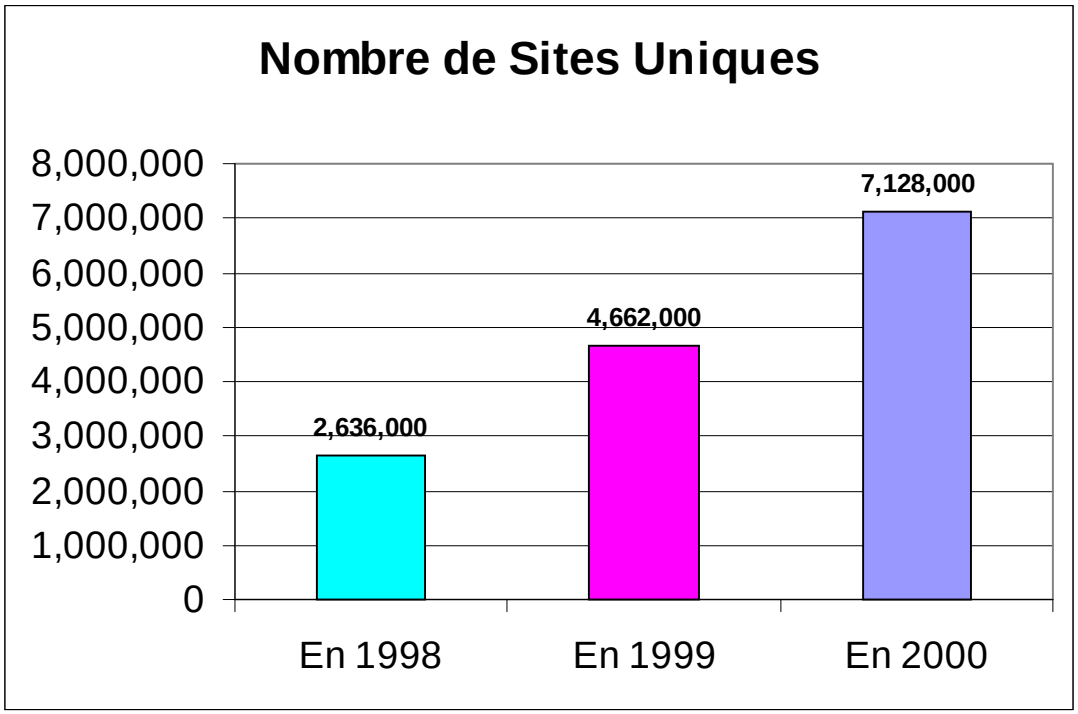


Figure 3

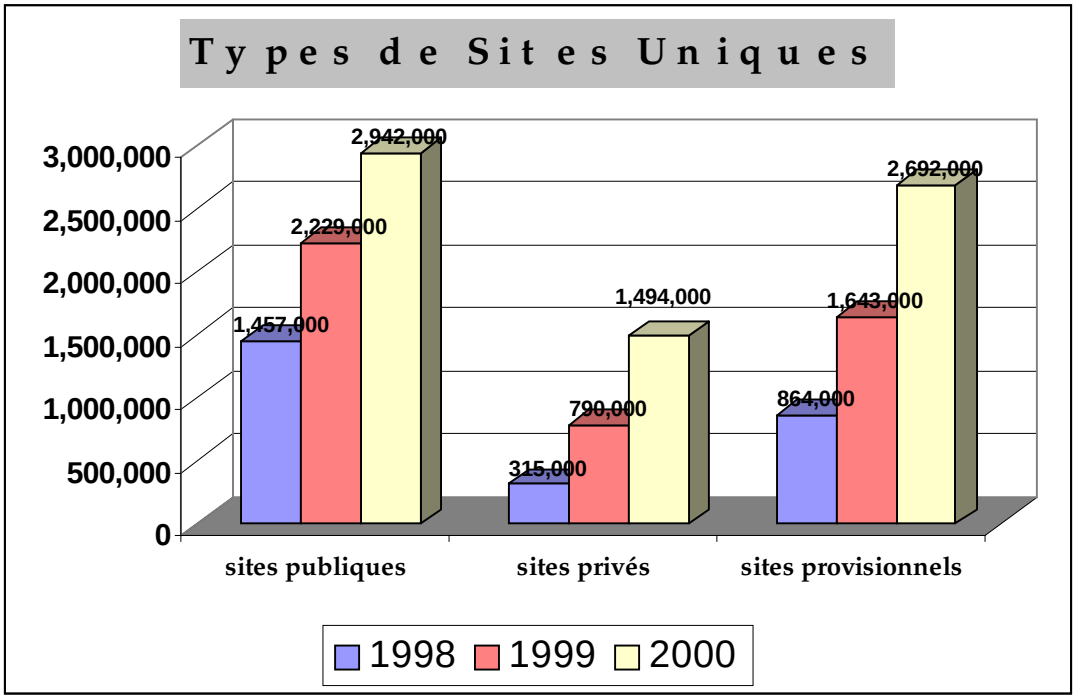


Figure 4

II-1-3 L'Etude de Cyveillance – juillet 2000 [110]

La troisième étude qui nous intéresse n'émane pas cette fois-ci d'une société savante, mais de l'entreprise privée « Cyveillance ». Cette étude a été réalisée avec le progiciel NetSapien Technology, un outil de recherche et d'analyse basé sur l'intelligence artificielle et dont la société se sert pour livrer des services de veille technologique sur commande.

Les résultats de cette étude ont été rendus publics le 10 juillet dernier, c'est donc l'étude la plus récente concernant la taille du Web visible, en plus, elle a le mérite d'être la seule étude qui a fait des prévisions sur la taille du web jusqu'au février 2001.

Méthodologie

- NetSapien Technology a été utilisé dans la collecte de données dans le but de créer un échantillon assez important à partir duquel Cyveillance a construit un modèle dynamique du Web. Dynamique dans le sens qu'il va au-delà de fournir une « image figée » du Web à un moment donné pour surveiller la croissance de celui-ci et les taux d'accélération de cette croissance de manière continue.
- Pour estimer la taille du Web, le modèle construit analysait les données spécifiques associées aux liens présents dans les pages examinées. En répétant ce processus continuellement, le modèle était capable de calculer la fréquence selon laquelle des URL uniques étaient rencontrés, la première fois et les fois suivantes. Ainsi, en comptant les URL uniques on arrivait à estimer la taille du Web, et en examinant le changement dans la fréquence de rencontre de ces URL entre deux passages, on pouvait mesurer le taux de croissance. En estimant la taille du Web sur une base continue, le modèle peut mesurer le taux d'ajout de nouvelles pages et de nouveaux sites ainsi que le taux de suppression de pages et de sites. En plus, il est capable d'estimer le taux d'accélération ou de décélération de la croissance du Web durant une période donnée.

Résultats de l'Etude

- L'étude a conclu qu'en juillet 2000 il y avait 2,1 milliards de pages uniques
- que le taux d'ajout de pages uniques par jour était de 7,3 millions
- que la croissance du Web va en s'accélèrant indiquant que celui-ci n'en a pas encore atteint le plus haut degré, après lequel cette croissance va se stabiliser ou ralentir.
- que poursuivant ce rythme de croissance, le Web visible atteindra 3 milliards de pages vers fin octobre 2000 et 4 milliards en février 2001.

En examinant la Figure 5 on constate qu'entre décembre 1999 (1 milliard de pages) et juin 2000 (2 milliards de pages)– c'est à dire en 6 mois - le Web a doublé de taille, et qu'avec les prévisions de Cyveillance, il doublera de taille encore une fois entre juin 2000 et février 2001.

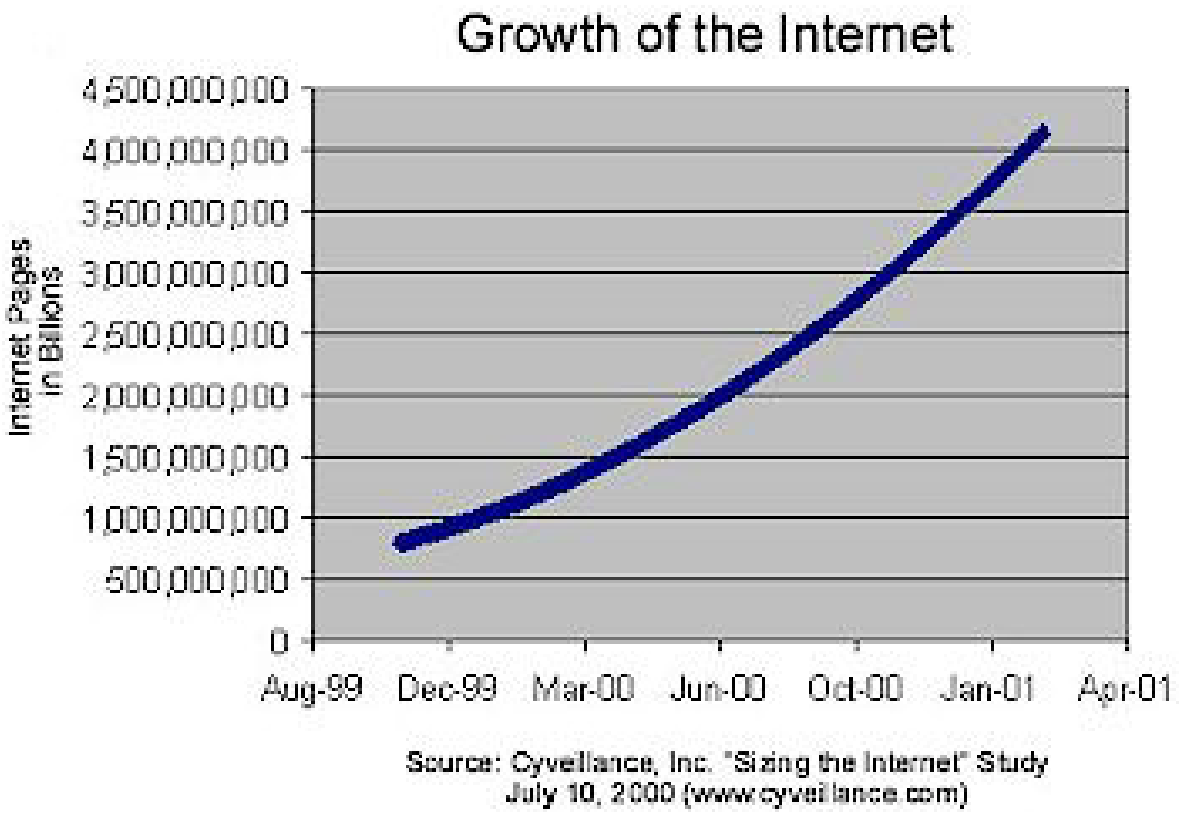


Figure 5 (copiée de l'étude de Cyveillance) [110]

Quelques caractéristiques du Web tirées de l'étude de Cyveillance	
La taille moyenne de la page	10.06 bytes
Nombre moyen de liens internes par page	23
Nombre moyen de liens externes par page	5,6
Nombre moyen d'images par page	14.38
Pourcentage des pages US contre les pages internationales	84,7% / 15.37%

Table 2 d'après Cyveillance [110]

II-2 Le Web Invisible ou "Deep Web"

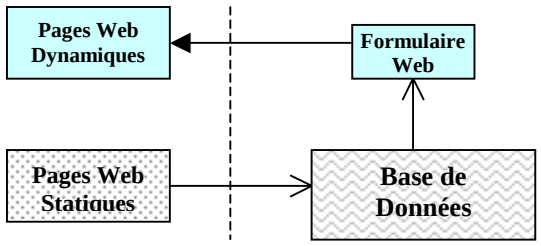
- On vise par Web Invisible[1], le web inaperçu par les robots des moteurs de recherche traditionnels durant leur parcours, et qui par ce fait est non-indexé, voire non-indexable, par ces moteurs. Ce sont les pages accessibles par mot de passe ou exigeant une autorisation préalable ou celles pour lesquelles il faut remplir un formulaire pour afficher les pages. Ce sont aussi les pages qui sont exclues de l'indexation par les moteurs de recherche utilisant les balises d'exclusion de robots. Plus important encore c'est toute l'information résidant dans les Bases de données

[1] L'expression "Invisible Web" fut lancée pour la première fois en 1994 par Dr. Jill Ellsworth pour designer l'information invisible aux moteurs de recherche traditionnels.

traditionnelles - qu'elles soient bibliographiques ou textuelles - et qui en réalité ne fait pas partie du Web mais qui, pour autant, a une interface Web.

- Il est faux donc de croire que cette information du fait qu'elle n'est pas indexable par les moteurs de recherche publics n'est pas indexée. Certes cette information est indexée mais par les établissements qui la produisent et la publient, c'est même l'information la plus organisée et la mieux indexée du Web. On comprend alors pourquoi dans les articles de Dr. Lawrence et Dr. Giles on désignait toujours le Web Visible par l'expression "Publicly Indexable Web" le mot "Publicly" juxtaposant toujours "Web Indexable", pour désigner l'information susceptible d'être indexée par les moteurs publics la différenciant ainsi de l'information indexée par ses producteurs privés. En fait, le dernier article de ces deux chercheurs[1] avait été mal interprété par les médias qui, en rapportant les chiffres de l'étude, ignoraient le mot "Publicly" et estimaient que le tout le Web comptait 800 millions de pages.**[42]**
- Contrairement au Web Visible qui lui est constitué de pages statiques, ce Web est constitué de pages Web dynamiques. Dynamiques?! Ce sont des pages qui n'existent en permanence sur le Web mais qui sont générées dynamiquement suite à une requête formulée. Alors que le contenu d'une page web statique est toujours le même, présentant la même information à tout le monde, (il a été crée manuellement par un web designer, placé sur un serveur web et est disponible à qui ou à quoi (un robot par exemple) visite le site, n'importe quel changement doit s'effectuer manuellement), une page web dynamiquement générée présente une information unique "personnalisée" pour les termes de la requête de l'interrogateur, seules les notices comportant tous les termes de la recherche et répondant aux critères requis seront affichées; le web est ici un moyen d'accès aux bases interrogées et un moyen de visualisation des résultats, après l'interrogation ces documents sont rapatriées dans la base.[2] [91]
- **Figure 6** explique le mécanisme de génération des pages dynamiques. Ces pages sont créées par un ordinateur utilisant un "langage" (souvent du CGI, Java ou Perl). Ce "langage" joue le rôle d'intermédiaire entre la demande de l'utilisateur soumise dans une page web statique (en pointillés) et entre une base de données où est stockée l'information (en vagues), le langage insère les résultats dans un formulaire et le présente à l'utilisateur à travers une page web générée dynamiquement. **[56]**

Figure 6 Génération d'une page web dynamique – Schéma traduit et reproduit d'après [56]



[1] L'article intitulé "Accessibility of the Information on the Web" **[61]**
[2] "They are never on the Web only the search box is." **[91]**

II-2-1 Le Web Invisible: début et explosion

On pourrait se demander à quelle date ces bases de données, qui existaient bien avant la création du World Wide Web et qui utilisaient d'autres protocoles Internet, ont commencé à basculer sur le web et à utiliser le protocole HTTP. La réponse à cette question implique un flash back dans l'histoire du W3. Au tout début du Web, il y avait un petit nombre de documents et de sites, toutes les pages ainsi conçues en HTML étaient constamment présentes et disponibles et par conséquent pouvaient être parcourues et indexées par les moteurs de recherche traditionnels; il suffit de dire qu'en juillet 1994 Lycos tenait la première place parmi les moteurs de recherche avec un indexe de seulement 54.000 documents.

Or, autour de 1996, trois phénomènes ont eu lieu: le premier c'était l'intégration de la technologie des Bases de données au web à travers des vendeurs comme Bluestone's Sapphire/Web et ensuite Oracle et les autres; le deuxième c'était la commercialisation du web initialement via les annuaires et les moteurs de recherche qui a rapidement évolué vers le e-commerce; le troisième phénomène était la révolution apportée aux serveurs pour permettre la présentation dynamique des pages web (cf. la technologie ASP de Microsoft et PHP de Unix). Ce confluent a fait du web une application orientée essentiellement vers les BDD, surtout pour les plus grands sites. Ainsi les gros producteurs d'information comme "Census Bureau", "Securities and Exchange Commission and Patents and Tradmarks Office", sans mentionner toutes les nouvelles entreprises nées directement sur Internet (Internet-based companies) voient dans le web le média idéal pour le commerce et le transfert d'information. **[92]**

II-2-2 L'Etude de BrightPlanet

Vu le volume considérable du Web Invisible, sa structure hétérogène et distribuée, aucune étude scientifique n'a été menée pour estimer sa taille ; on comptera uniquement un nombre d'articles scientifiques (ou de presse) définissant ce Web et précisant son étendue pour sensibiliser les utilisateurs aux richesses qui y sont contenues, ou des sites fédérateurs et spécialisés compilant des ressources qui en font partie. Il existe cependant une initiative très intéressante pour estimer la taille du Web Invisible, c'est celle de la société BrightPlanet qui développe le site "CompletePlanet.com", recensant 20.000 ressources du Web Invisible, ainsi qu'une technologie de recherche de ce Web connue sous le nom de "LexiBot", un métamoteur offline capable de rechercher et d'interroger les ressources du Web Invisible.

On notera que BrightPlanet préfère utiliser les termes de "Deep Web", qu'elle oppose à surface Web, plutôt que ceux de Web invisible et visible, couramment employés, **[53]** car pour elle le problème n'est pas un problème de visibilité ou d'invisibilité, mais de la technologie de recherche utilisée par les moteurs classiques qui est incapable d'explorer les ressources du Web Invisible, alors que d'autres outils comme les "directed search engines" eux sont capables d'aller dans les profondeurs du Web et d'explorer ses richesses. **[92]**

Méthodologie [92]

- Pour pouvoir comparer Deep et Surface Web sur une base commune, BrightPlanet a créé un formulaire HTML vers lequel elle expédiait les notices récupérées dans les BDD suite à une interrogation par LexiBot.[1]
- Ce logiciel, défini comme un "directed query engine", génère simultanément un grand nombre de requêtes pour interroger différentes BDD en même temps chacune selon ses particularités.
- Les documents obtenus en retour, sont téléchargés dans une BDD, où ils vont ensuite subir un nombre de traitements - toujours avec ce logiciel -: ils sont tout d'abord indexés, ensuite triés suivant leur pertinence en utilisant quatre différents algorithmes de classement.
- Parmi un total de 100.000 sites du Deep Web, BrightPlanet a choisi aléatoirement un échantillon de 17.000 pour constituer son champs d'étude. Sa méthode consistait en ce qui suit:
  - estimer le nombre total de documents par site
  - récupérer au moins 10 résultats dans chaque site et calculer la taille moyenne du document en octets tout en incluant les balises HTML
  - multiplier cette taille par le nombre de documents contenus dans ce site pour avoir le volume total du site

Le comptage des notices:

Le nombre total de notices par site a été obtenu par différents moyens:

- envois d'emails aux webmasters des sites testés pour vérifier le nombre de notices ainsi que le volume de l'information stockée
- le nombre de notices, rapporté par les sites eux-mêmes
- la taille des sites telle qu'elle a été annoncée lors de certaines conférences
- le recours à la fonction de recherche de certains sites qui en réponse à une requête combinant l'opérateur booléen d'exclusion "NOT" et un terme imaginaire (qui n'existe pas dans la base), donne le nombre absolu de notices

Résultats de l'Etude [92]

- Le "Deep Web" contient environ 550 milliards de documents uniques
- le nombre total de sites du web invisible dépasse les 100.000
- les 60 sites les plus importants représentent à eux seuls plus de 40 fois le volume du web visible[2]

[1] Dans le stockage actuel des données du Deep Web, le volume de l'information est beaucoup plus inférieur aux chiffres qu'on trouvera ici, [92] car la méthode employée compte aussi les balises HTML du formulaire créé dynamiquement pour fournir un dénominateur commun de comparaison entre Deep et Surface Web.

[2] Notons que les résultats de cette étude sont comparés aux chiffres concernant le Web visible (estimé à 1 milliard de documents), tirés du communiqué de presse publié par Inktomi en janvier

- le Deep Web représente la catégorie croissant le plus rapidement sur Internet
- 95% du Web Invisible est accessible gratuitement sans restriction ni inscription

Figure 7 illustre la distribution des sites du Web Invisible par domaines, et Figure 8 illustre la distribution de ces sites par types de contenu. On en constate la valeur ajoutée de ce Web, dominé principalement par les Bases de données (Banques de données + sites internes) et les journaux électroniques (publications), sources inestimables de l'information scientifique sur le Web.

Alors que les plus vastes sites du Deep Web sont largement connus par les internautes, le site typique de ce Web (les sites représentant la valeur médiane en taille et en volume) est méconnu.

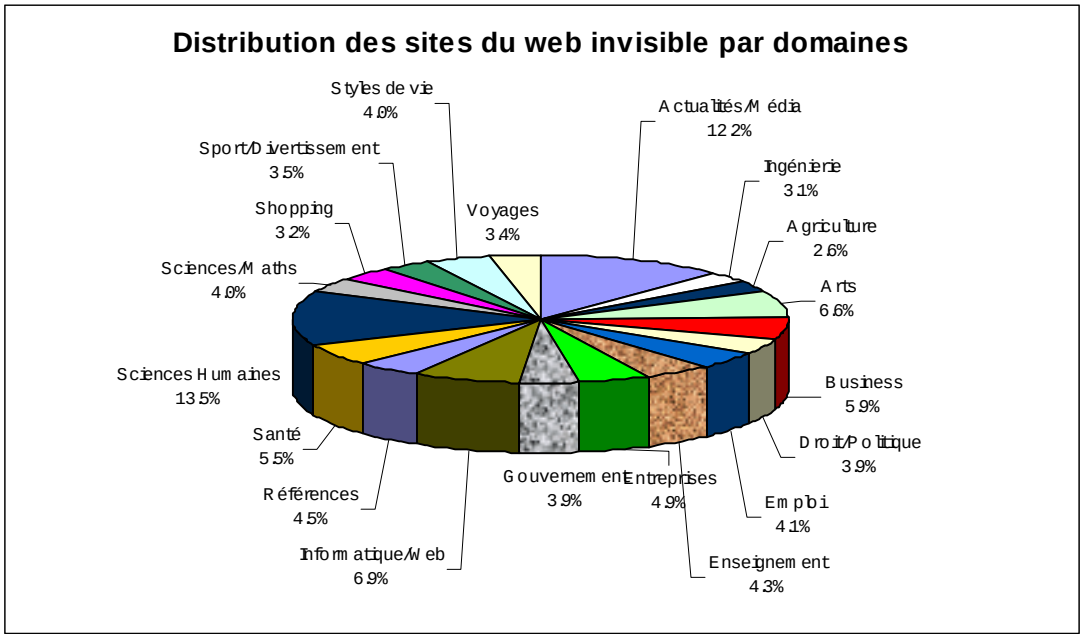


Figure 7: conçues d'après les chiffres de BrightPlanet

2000; les tests du web visible et ceux du web invisible ont donc été effectués à plusieurs mois d'écart, ce qui fausse en quelque sorte la comparaison. Nous essayerons plus loin de comparer les chiffres de la présente étude à ceux de l'étude de Cyveillance qui a été publiée en même temps pour le bien fondé de la comparaison.

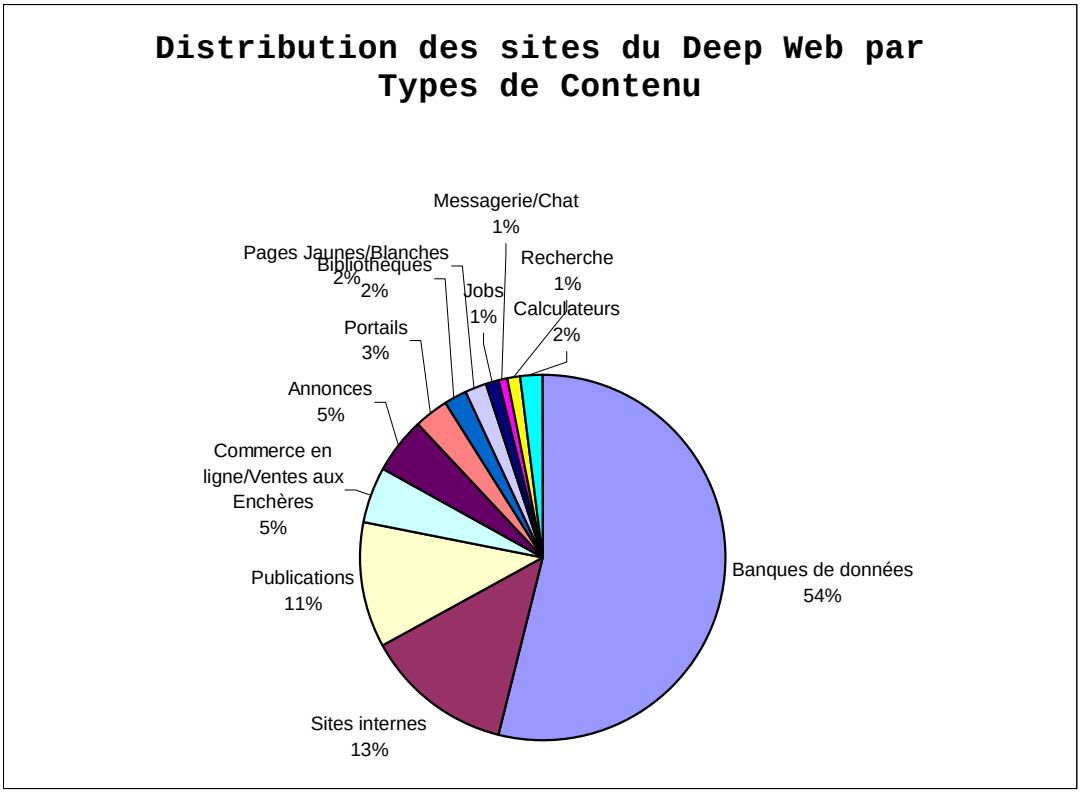


Figure 8: traduite et reproduite d'après la figure de BrightPlanet

II-3 Le Web Visible comparé au Web Invisible [92] [110]

NB:
Dans l'étude de BrightPlanet tous les chiffres sont comparés aux chiffres du communiqué de presse d'Inktomi publié en janvier 2000 qui estimait la taille du Web Visible à un milliard de pages. Cependant, pour plus d'exactitude et pour rendre pertinente la comparaison, nous allons comparer ces chiffres à ceux de l'étude de Cyveillance qui elle estimait l'existance de 2.1 milliards de pages Web en juillet 2000. On notera alors que certains des chiffres figurant ci-dessous, ont été obtenus en divisant par 2 les estimations de BrightPlanet, d'autres ont été directement copiés de l'étude de Cyveillance. On marquera par un * les chiffres qui ont été modifiés pour les distinguer de ceux de BrightPlanet.

II-3-1 De la Taille

- Le Web Invisible est 275* fois plus large que le Web Visible.
- les 60 plus larges sites du Web Invisible représentent à eux-seuls plus de 20* fois le volume du Web Visible
- le Web Invisible contient 7.440 tera octets d'information contre 21* tera octets pour le Web Visible[1]

[1] 1 tera octets = 10^3 Go = 10^6 Mo = 10^9 Ko

II-3-2 Du Taux de Croissance et de la Fraîcheur de l'Information

Le Web Invisible croit beaucoup plus rapidement que le Web Visible: des notices sont ajoutées régulièrement aux différentes BDD et aux OPAC, de nouveaux numéros de journaux électroniques ou de la presse publiée sur le Web paraissent tous les jours, toutes les semaines, etc… selon leur périodicité, - les anciens numéros restant consultables parce qu'archivés -, tout ceci en plus des nombreuses bibliothèques qui de plus en plus rendent leurs catalogues accessibles sur le Web. A l'opposé, les pages d'un site web visible donné une fois placées sur le web leur contenu ne bouge pas beaucoup.

II-3-3 De la Duplication et de l'Unicité du Contenu

La duplication est un phénomène propre au Web Visible; certains sites ont au moins un site miroir qui reproduit le même contenu sur une autre adresse IP, les grands annuaires ont même plusieurs sites miroirs (cf. Yahoo! Altavista, etc….). Le fait aussi que le contenu du Web Visible est accessible à tout le monde fait que certains s'approprient les documents les plus populaires et les placent sur leurs propres sites; l'information commune comme les communiqués de presse, les logiciels et les listes de produits paraissent maintes fois dans les listing des moteurs de recherche et les moteurs eux-mêmes dupliquent l'information. Bref, la duplication est fonction de la disponibilité et de l'accessibilité des documents et est une nécessité dictée par le principe du marché qui régit le Web. Du fait qu'il n'est pas facile à découvrir, le contenu du Web Invisible n'est pas – du moins aisément - copié par des tiers et reproduit sur d'autres sites. Toutefois, des duplications ont été observées parmi les catégories qui présentent des informations très importantes et largement consultées comme les pages jaunes/blanches, les registres généalogiques, …, ces catégories représentent moins de 20%[1] de l'ensemble et elles sont toujours dupliquées par leurs producteurs.

II-3-4 De la Qualité

Le contenu du Web Invisible est strictement contrôlé car validé par des scientifiques, soumis à des comités de lecture avant la publication, critiqué, et l'origine de la plupart des sites est clairement identifiée. [53] Vue la facilité de la publication sur le Web visible, n'importe quelle personne peut créer un site avec certains nombres de pages et le placer sur le web sans passer par des tiers pour juger de son contenu. Il est donc clair que l'information obtenue à partir du Web Invisible est d'une qualité beaucoup supérieure à celle obtenue du Web Visible, BrightPlanet multiplie par "2000" la qualité de cette information si on emploie un seul moteur de recherche et par "400" pour une recherche combinant plusieurs moteurs ou employant un métamoteur.

II-4 Typologie du Web: Comptage des Domaines "DNS"

Internet Domain Survery de Network Wizards effectue des relevés semestriels (janvier, juillet) de l'Internet mondial. Les relevés de Network Wizards prennent la suite de ceux du

[1] Les 80% restant constitués de banques de données, de publications, d'indexes internes de certains sites et de catalogues de bibliothèques présentent un contenu unique non doublé. [92]

SRI pour constituer une série continue depuis août 1981 (fréquences variables cependant). Il présente un décompte par TLD (Top Level Domain) pour l'ensemble de l'Internet.

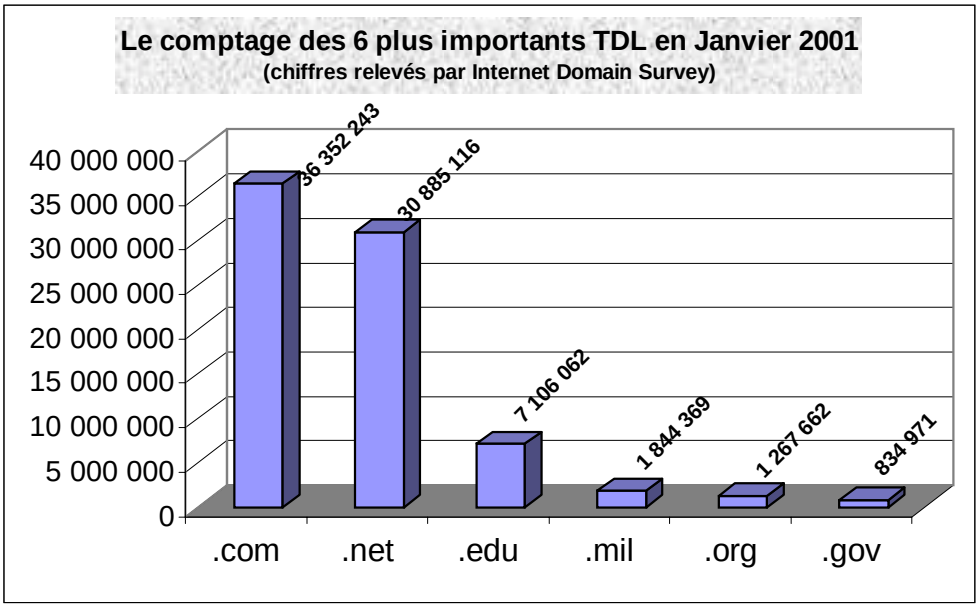


Figure 9 Conçue d'après les chiffres de [101]

com = le domaine commercial, **net** = les réseaux, **edu** = éducation,
mil =militaire, **org** = organisations, **gov** = gouvernemental

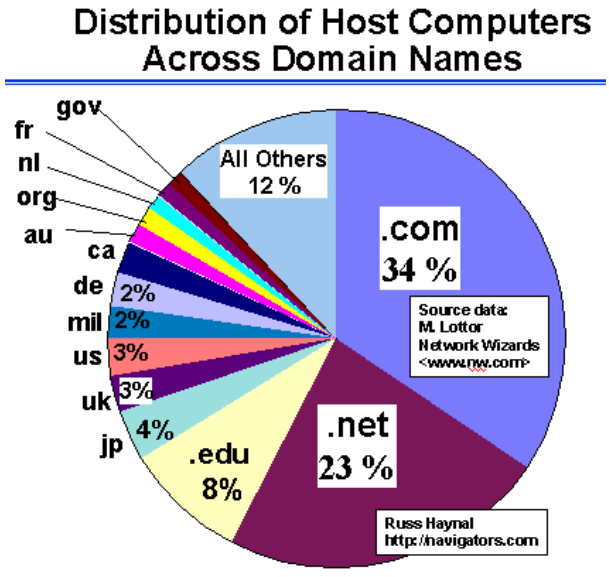


Figure 10 Copiée de [100]

Examinant la Figure 9et la Figure 10, on constate que la grande majorité des sites est partagée entre le TLD "com" et le TLD "net" (environ 57% des hôtes).

La langue prédominante du Web est l'anglais (environ 80% des sites), vient ensuite l'allemand puis le français puis l'italien.

Le Comptage des Domaines est-il Représentant de l'Offre?

- Pas nécessairement, car certains hôtes n'ont pas de contenu sur leurs pages d'accueil (cf. les FAI)
- Beaucoup d'hôtes référencés dans le DNS n'existent pas physiquement ou ne sont pas en service.

II-5 L'Estimation de la Taille du Web: Pourquoi?

Après avoir fait la topologie du Web, estimé sa taille et sa croissance et tracé la ligne de démarcation entre Web Visible et Web Invisible, une question s'avère pertinente: à quoi sert tout ceci? pourquoi tant de temps, d'effort et d'investissements aussi bien de la part de sociétés savantes (comme le NEC research institute et le bureau de recherche de l'OCLC) que de celle de sites commerciaux (comme Cyveillance, Inktomi et BrightPlanet) consacrés à cette question? L'objectif serait-il si naïvement exprimé: au dixième anniversaire du Web et parti de zéro on est arrivé à tel nombre de sites, tel nombre de pages et tel nombre de serveurs web?! Que de dépenses, de temps passé, d'efforts fournis pour un aboutissement qui ne vaut pas le coup!!!

Or, si le but de sites comme Cyveillance ou BrightPlanet peut être purement commercial, visant à faire prévaloir telle technologie de recherche ou à commercialiser tel produit auprès des entreprises qui cherchent un outil puissant de veille technologique ou d'intelligence artificielle, le but des sociétés savantes et des chercheurs est tout autre, c'est même une complexité d'objectifs. Pour les uns le but de telle démarche est de faire le bilan entre le volume de l'information scientifique sur le web et l'accessibilité de cette information aux scientifiques via les moteurs de recherche **[61] [64]**; pour les autres, c'est tâtonner le cyberespace avant d'investir davantage dans le catalogage des ressources Internet **[105]**; pour les troisièmes c'est une mesure de prudence pour ne pas adhérer sans réserve aux opinions qui voient dans le web la source ultime d'information pour le 21^{e} siècle. **[87]**

II-6 Le World Wide Web et l'Information: quelle relation?

Autour de la question du web comme source d'information, les opinions sont partagées, surtout dans les milieux des bibliothécaires dont le traitement de l'information constitue le cœur du métier. Une question est soulevée: faut-il cataloguer les ressources du web? Quelques bibliothécaires fanatiques de l'imprimé voient que le contenu du web est nul, éphémère, et que pour ces deux raisons comme pour celle que les techniques de catalogage ont été conçues uniquement pour l'imprimé et ne sont pas applicables au web[1], il ne faut pas s'engager dans une telle entreprise.[83] Cette opinion est très extrémiste et elle est contrariée par la pratique des bibliothèques publiques et universitaires aux Etats-Unis et dans les pays occidentaux, qui de plus en plus fournissent aux usagers un accès à l'Internet, et dont les personnels mettent beaucoup de temps à compiler sous forme de

[1] "Web content is trash, Web content is too ephemeral, and cataloging technologies were designed for print and are not applicable to the Web"

signets ou autres des ressources électroniques. Si on compte aussi les frais élevés d'accès et de connexion, l'équipement de fonctionnement, la maintenance, le personnel qualifié pour assurer ce service, on peut imaginer que des sommes considérables ont été investies à moyen et à long termes là-dedans et que ces dépenses ne sont pas sans justification.

Non seulement l'Internet a envahi les salles de lecture dans les bibliothèques, mais aussi les bureaux des bibliothécaires et des catalogueurs qui y puisent des sources inestimables d'information spécialisée et des outils de travail.

Dans leur excellent article[1], Dr. Lawrence et Dr. Giles soulignent l'infiltration du Web dans tous les domaines de la vie: depuis l'achat de biens et la localisation de services jusqu'à l'accès à l'information scientifique. A ce sujet, ils ne manquent pas de mettre l'accent sur l'importance primordiale du Web à la diffusion de l'information parmi les savants et les chercheurs. Ils centrent leur intérêt sur le Web indexable par les moteurs de recherche et concluent "qu'environ 6% des serveurs Web hébergent un contenu éducatif ou scientifique (cf. les serveurs d'universités, de collèges et des laboratoires de recherche), que le Web contient différentes sortes de ces matériaux scientifiques, à savoir les pages personnels des savants et des chercheurs, des preprints, des rapports techniques, des actes de conférences, des journaux, des matériaux d'enseignement, et que la grande majorité de cette information n'existe pas dans les BDD traditionnelles."[2] **[61]**

Ceci explique le grand intérêt qu'une organisation comme l'OCLC a porté au Web depuis sa création, soit en lui consacrant NetFirst, l'une des bases de la famille FirstSearch, pour cataloguer t et indexer ses ressources, **[83]** soit en développant et en mettant à la disposition des utilisateurs et des concepteurs de sites un outil de description de sites et de pages à leur portée: le Dublin Core Metadata Initiative(DCMI), qui est une sorte de CIP (Cataloging In Print) mais pour les ressources électroniques. Enfin, l'OCLC a couronné ces initiatives par le projet CORC, qui combine les efforts de 489 bibliothèques dans 24 pays, dans le but de construire une base de données pour les pages Web dont le contenu est d'une très grande utilité pour les bibliothèques. Ainsi CORC avec le WorldCat constitueront une base extrêmement riche et normalisée à une échelle internationale, qui mettra à la disposition des usagers des bibliothèques du monde entier, le matériel unique de chaque partenaire du projet. **[81]**

II-7 L'Accessibilité de l'Information sur le Web

Les moteurs de recherche restent le moyen essentiel d'accès à l'information sur le Web; 85% des internautes y recourent pour retrouver ce qu'ils cherchent. [61] Des sites spécialisés comme "Search Engine Watch" et "Search Engine Show Down" s'intéressent uniquement à évaluer la performance de ces moteurs de recherche, à comparer la taille de leurs indexes et à analyser leurs méthodologies d'indexation et de recouvrement de l'information, leur méthode de tri des résultats fournis, tout ceci dans le but de mesurer leur couverture du Web et la pertinence de l'information qu'ils présentent aux utilisateurs.

---

[1] Accessibility of the Information on the Web

[2] Il rappèlent que tous les articles scientifiques et de recherche qui ne sont accessibles maintenant qu'à travers les grandes BDD payantes ont tout d'abord existé sur le web.

II-7-1 Performance des Moteurs de Recherche

- En décembre 1997, utilisant 575 requêtes formulées par le personnel du NEC Research Institute pour interroger les 6 gros moteurs de l'époque Dr. Lawrence et Dr. Giles arrivent à la conclusion que le moteur le plus important n'indexait qu'un tiers du Web (HotBot 34%, suivi d'AltaVista: 28%) et que la couverture combinée des six moteurs couvrait environ 60% du Web Visible. [63] [108]
- En février 1999, à peine plus d'un an après, ces chercheurs recommencent l'expérience en augmentant le nombre de moteurs à 11 et en multipliant par deux le nombre des requêtes (1050 requêtes). Les résultats obtenus sont très décevants. Le moteur ayant la couverture la plus importante est désormais NorthernLight qui indexe 16% seulement du Web. La couverture combinée des 11 moteurs représente quant à elle 42% du nombre total de pages Web!!!! **[61] [109]**
- Les résultats obtenus montrent que les moteurs de recherche prennent de plus en plus de retard face à la croissance exponentielle du Web. **[61]**
- Un second problème est celui de la mise à jour des indexes des moteurs. Les auteurs ont constaté que l'indexe le plus exhaustif est le moins rafraîchi et vice-versa; le pourcentage de liens invalides de pages inexistantes ou qui ont changé de URL est aussi inquiétant[1]; l'âge moyen des pages stockées dans les indexes des robots est de 186 jours tandis que l'âge médian est de 57 jours. **[61] [109]**
- On imagine alors la frustration des utilisateurs dont la seule source d'accès à l'information enligne sont les moteurs de recherche: à aucun moment ces utilisateurs n'ont une image fidèle du Web. En fait, les internautes qui recherchent le Web à travers les moteurs de recherche n'ont pas accès à l'information existante sur celui-là, mais, - à leur grand étonnement - aux pages stockées dans les indexes des moteurs de recherche et qui peuvent être à plusieurs mois de décalage avec ce qui se trouve effectivement et réellement sur le Web.

Pourquoi les moteurs de recherche n'indexent qu'une petite fraction du web? Il existe un point au-delà duquel il n'intéresse plus aux moteurs d'améliorer leur performance, car ça devient lourd pour eux en matière d'économie et de finances. La gestion d'un indexe volumineux nécessite des logiciels très puissants, des équipements plus performantes, des coûts de maintenance importants et implique souvent des temps de réponse plus longs. **[61] [48]**

II-7-2 Frustration de Part et d'Autre

L'une des grandes promesses du Web fut l'égalité d'accès à l'information, en d'autres termes la démocratisation de diffusion de toute sorte d'information sur le réseau. Notre droit comme utilisateurs et internautes est de demander si cette promesse a été remplie, sinon c'est la responsabilité de qui? des moteurs de recherche qui dans la plupart du temps

[1] Les nouvelles pages ajoutées ou modifiées juste après le passage du robot peuvent attendre des mois avant d'être indexées lors du suivant passage, en plus certains sites présentent des pages faites exprès pour les robots dont le contenu est totalement différent des pages auxquelles on a accès si on visite la même URL.

sont notre seul moyen d'accès à l'information? ou bien les producteurs de cette information?

Malheureusement la frustration est de part et d'autre.

1.1.1.9. Du côté des moteurs de recherche

- Les moteurs de recherche ont des partis pris dans le choix de "l'échantillon du Web" à indexer. Parce que les moteurs suivent les liens à la recherche de nouvelles pages pour enrichir leurs indexes, ils ont tendance à rechercher et indexer les pages vers lesquelles pointent plus de liens. **[61]** Or si on sait que la distribution de l'information sur le Web ressemble plutôt à un nœud de papillon qu'à une toile d'araignée, (Figure 11) on constatera que les pages les mieux servies sont ceux situées au centre du réseau. **[95] [52]**
- Certains moteurs comme Google et DirectHit s'appuient sur la popularité des pages pour trier les documents selon l'ordre de pertinence. Ces techniques non basées sur la description des ressources documentaires défavorisent l'accessibilité à l'information. Ceci contribue à l'accroissement de la popularité des pages les plus populaires pendant que les nouvelles pages ne bénéficiant pas de liens pointant vers elles éprouvent des difficultés de plus en plus grandissantes pour avoir leur place dans le listing des moteurs de recherche. Ceci peut retarder voire même constituer un obstacle intransigeant devant la diffusion et la visibilité d'une nouvelle information de très haute qualité. **[61]**

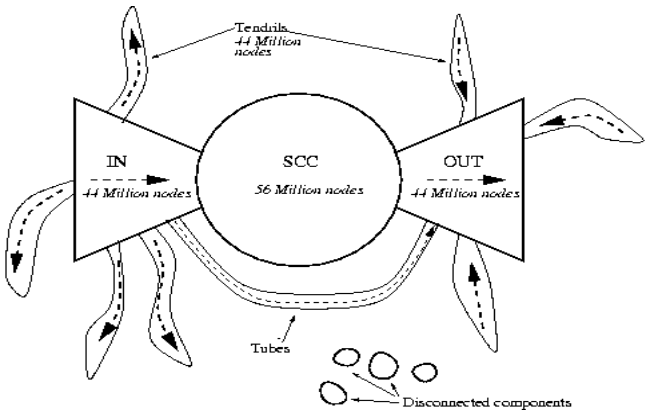


Figure 11: copiée d'une étude réalisée par les centres de recherche d'IBM, de Compaq et d'Altavista. [95]

Explication de la figure:(copiée de [52])
Plus de 90% du Web forme un ensemble de pages connectées entre elles mais de façon inégales. Ces pages sont réparties en 4 grands ensembles:

- *la partie centrale (SCC: strongly connected component) est constituée du noyau ultra connecté, et contient 28% des pages Web. La navigation y est aisée. Ce noyau compact constitue le cœur du réseau, c'est lui qui permet de passer, par clics successifs, de n'importe quelle page du IN vers une page du OUT. Ce sont les pages du cœur que les robots des moteurs de recherche indexent en priorité, et c'est à partir de leurs liens qu'ils explorent le Web.*

- *la partie gauche (IN) contient les pages "d'origine" et représente 21% du réseau. Ses pages offrent des liens vers le cœur du Web, mais l'inverse n'est pas vrai. Ces pages peuvent être reliées entre elles, mais il n'y a pas de liens issus du cœur qui pointent vers elles.*
- *la partie droite (OUT) correspond aux pages de "destination"; elles représentent également 1/5 du réseau. Ces pages sont accessibles depuis le cœur du Web, mais aucun retour n'est possible.*
- *une dernière zone représentant également 1/5 du Web (TENDRILS), est composée non connectées au cœur du réseau. Ces pages sont accessibles depuis les pages d'origine et/ou donnent accès aux pages de destination.*
- *enfin, près de 10% des pages Web sont totalement déconnectées des autres pages.*

1.1.1.10.Du Côté des Producteurs

- Nos chercheurs (comme beaucoup d'auteurs d'ailleurs) se sont intéressés à analyser les métadonnées dans les pages d'accueil de chaque serveur Web en étudiant les méta balises HTML.[1] Une variété de méta balises est utilisée dont la plupart encodent des détails qui ne servent pas à identifier le contenu de la page. Les auteurs ont cependant considéré la présence d'un ou de plusieurs mots-clés ou de balises de description comme un effort qui a été fait en vue de la description du contenu du site. **[61]**
- Ils ont constaté que seulement 34.2% des serveurs contiennent des métadonnées de ce types dans leurs pages d'accueil. L'usage relativement bas des simples métadonnées HTML suggère que l'acceptation et l'utilisation généralisée de standards plus complexes et plus sophistiqués comme XML ou le Dublin Core, seront très lentes (0.3% des sites du Web Indexables contenaient des métadonnées basées sur le Dublin Core).
- Les auteurs ont en outre remarqué une diversité très énorme dans les méta balises HTML utilisées, avec 123 balises distinctes. On en constate qu'on est encore loin d'une normalisation.
- Ajoutons à tout ceci l'usage abusé des métadonnées par certains sites: il y a des sites qui, pour bénéficier d'un meilleur classement par les moteurs, utilisent des métadonnées qui ne reflètent pas vraiment le contenu des pages décrites; les moteurs étant au courant de ces pratiques, la plupart d'entre eux ignorent ces données, lesquelles ,devaient normalement les aider à indexer les sites pertinemment. **[81]**

[1] Ces médatadonnées servent à définir le contenu intellectuel des sites par les concepteurs de ces sites eux-mêmes.

L'Avenir de la Recherche de l'Information sur le Web

Faut-il conclure à l'échec – ou du moins la défaillance – des outils de recherche face à un Web à une croissance exponentielle? Et quelle sera la situation si on lance Internet 2 avec la nouvelle version d'adresses IP qui multiplieront à l'infini les hôtes et les sites, et que le Web sera complètement débridé? Pour l'instant, ces questions semblent sans réponses. Toutefois, il ne faut pas porter des jugements hâtifs, car depuis ces études, les moteurs ont amélioré considérablement leur performance et élargi leurs indexes. Andrei Broder, chef scientifique à Altavista, affirme que la capacité du matériel informatique et l'efficacité des logiciels vont de pair avec la croissance du réseau **[86]**; et avec le développement du Web sémantique et la nouvelle génération des moteurs capables de traiter le langage humain, il y a de quoi à être optimistes.

Il ne faut pas aussi oublier que les moteurs de recherche ne sont pas notre seule fenêtre sur le Web. Les annuaires, les sites fédérateurs, les portails, les vortails (portails verticaux), et tous ces outils qui font appel à l'expertise humaine constituent des moyens organisés et structurés d'accès à l'information sur le Web. Les moteurs eux-mêmes tendent à la "portalisation"[1]. Tout ceci, en sus des efforts des chercheurs et des savants pour rendre visibles et accessibles aux utilisateurs du réseau les ressources de l'Internet: Dr. Steve Lawrence est parti à la conférence de NFAIS 2001, qui s'est tenue aux Etats-Unis, du 25 au 28 février dernier, non pas pour parler toujours de la taille du Web, mais pour étudier avec ses homologues la constitution d'un Immense Portail pour l'Information Scientifique sur le réseau.

Dernier mot:

"With anything as dynamic and rapidly expanding as the Web, any measurement is a snapshot frozen in time."

Michael Dahn

[1] "Portalization" est un terme qui fut employé par Danny Sullivan du Search Engine Watch lors de la quatrième conférence sur les moteurs de recherche, pour désigner la tendance des moteurs à devenir des portails. L'objectif selon lui est purement commercial visant à garder l'utilisateur sur le même site en l'attirant par le biais du mail, du chatting ou du shopping enligne, après avoir échoué à satisfaire sa requête. **[20]**

III- BIBLIOGRAPHIE

III- BIBLIOGRAPHIE

III-1 Actes de congrès et de conférences

[1] **BERGMAN Michael.** Content discovery and aggregation on the Web. In *Networking @ Internet Speed, Keynotes on Information Strategy, Technology and Vision, NFAIS 2001, Philadelphia, Pennsylvania, 25-28 February 2001.* [Online]. Available from Internet: <URL: http://www.pa.utulsa.edu/nfais/Conf2001/Conf2001.final_prog.html>

[2] **BHARAT K., BRODER A.** Mirror, mirror on the Web: A study of host pairs with replicated content. In *The WWW8: 8th International World Wide Web Conference, Toronto, Ontario, 11-14 May 1999. Computer Networks - The International Journal Of Computer And Telecommunications Networkin*, Amesterdam : Elsevier, vol. 31, no. 11, 1999. p 1579-1590.

[3] **BHARAT Krishna, BRODER Andrei.** A technique for measuring the relative size and overlap of public Web search engines. In *7th International World Wide Web Conference, Australia, Brisbane, 14-18 April 1998.* [Online]. Available from the Internet: <URL: http://www-sor.inria.fr/mirrors/www7/programme/fullpapers/1937/com1937.htm>
Study update: <URL: http://research.compaq.com/SRC/whatsnew/sem.html>

[4] **BRAY Tim.** Measuring the Web. In *5th International World Wide Web Conference, Paris, 6-10 May 1996.* [Online]. Available from Internet: <URL: http://www5conf.inria.fr/fich_html/papers/P9/Overview.html>

[5] **BREWINGTON Brian E., CYBENKO George.** How dynamic is the Web? In *Ninth International World Wide Web Conference, Amesterdam, 15-19 May 2000. Computer Networks - The International Journal Of Computer And Telecommunications Networking,* Amesterdam: Elsevier, June 2000, vol.33, no.1-6, p.257-76. Also available from Internet: <URL: http://www9.org/w9cdrom/index.html>

[6] **BRUNNING D.** How are you going to find it: what libraries don't know, think they know, want to know, and should know about Web search engines. In *Proceedings of 21st Annual National Online Meeting, New York, U.S.A, 16-18 May 2000.* Edited by M.E. Williams. NJ: Information Today, 2000. p.55-60.

[7] **GILES C. Lee.** Searching the Web (keynote address): can you find what you want? In *Proceedings of the eight international conference on Information knowledge management, Kansas City,2-6 November 1999, p.1-2.* [Online] Available from Internet: <URL: http://www.kric.ac.kr:8080/pubs/citations/proceedings/cikm/319950/p1-giles/#abstract>

[8] **HA W.-G., DOA HOOM KIM, JUHN J.** Dynamic Analysis of Internet Growth. In *Proceedings Of The 16th International Conference-System Dynamics Society, Quebec, July 7 1998.*

[9] **KISHI Nobuko, OHMORI Takahiro, SASAZUKA Seiji, …** Estimating Web Properties by Using Search Engines and Random Crawlers. In In *INET 2000 Proceedings: The Internet Society Conference, Yokohama, 18-21 July 2000.* [Online]. Available from Internet: <URL: http://www.isoc.org/inet2000/cdproceedings/2a/2a_3.htm>

[10] **LAVOIE Brian, O'NEILL Edward, McCLAIN Patrick.** Web characterization using sampling method. In *W3C Web characterization, 5 Nov. 1998.* OCLC Office of Research. [Online]. Available from Internet: <URL: http://wcp.oclc.org/>

[11] **LAWRENCE Steve.** Information Portals: The challenge of new entrants. In *Networking @ Internet Speed, Keynotes on Information Strategy, Technology and Vision, NFAIS 2001, Philadelphia, Pennsylvania, 25-28 February 2001.* [Online]. Available from Internet: <URL: http://www.pa.utulsa.edu/nfais/Conf2001/Conf2001.final_prog.html>

[12] **LAWRENCE S., GILES C. L.** Searching the Web: general and scientific information access. In *Proceedings of 1st IEEE-RPS Joint Conference on Internet Technologies and Services, Moscow, Russia, 25-28 October 1999.* NJ: IEEE, Piscataway, 1999. p. 18-31.

[13] **LAWRENCE Steve, BOLLACKER Kurt, GILES Lee.** Indexing and retrieval of scientific literature. In *Proceedings of the eight international conference on Information knowledge management, Kansas City,2-6 November 1999, p.139-146.* [Online] Available from Internet: <URL: http://www.kric.ac.kr:8080/pubs/citations/proceedings/cikm/319950/p139-lawrence/#abstract>

[14] **O'NEILL Ed.** Characteristic of Web accessible information. A presentation given at *1997 IFLA Conference, Copenhagen, 2 Sept. 1997.* Office of Research OCLC Online Computer Library Center, Dublin, Ohio USA. IFLA Journal, 1998, no. 24, p.114-116. [Online]. Available from Internet: <URL: http://www.oclc.org/oclc/man/ifla/index.htm>

[15] **PRICE Gary, SHERMAN Chris**. The Invisible Web. In *15th Annual AIIP (The Association of Independent Information Professionals) Conference, New Orleans, 19-21 April 2001.* [Online]. Available from Internet: <URL: http://www.aiip.org/conf2001program.html>

[16] **PRICE Gary, SHERMAN Chris**. The Invisible Web. In *InfoToday 2001:, The Global Conference and Exhibition on Electronic Information and Knowledge Management, New York, 15-17 May 2001.* [Online]. Available from Internet: <URL: http://www.infotoday.com/it2001/nationalonline.htm>

[17] **RISVIK Kunt Magne.** Scaling with the Web – search engine challenges. In *4th Annual Search Engine Meeting, Boston, 10-22 April 2000.* [Online]. Available from Internet: <URL: http://www.infonortics.com/searchengines/sh00/risvik_files/frame.htm>

[18] **SHERMAN Chris, PRICE Gary.** Searching the Invisible Web: Resources & concepts For Web Research. In *13th Annual Conference of the Association of Professional Researchers for Advancement, California, 26-29 July 2000.* Brochure de la Conference en Format PDF : <URL : http://www.aprahome.org/2000%20Conf%20Broch%20Final.pdf>

[19] **SHIODE Narushige, BATTY Michael**. Power law distribution in real and virtual Worlds. In *INET 2000 Proceedings: The Internet Society Conference, Yokohama, 18-21 July 2000.* [Online]. Available from Internet: <URL: http://www.isoc.org/inet2000/cdproceedings/2a/2a_2.htm>

[20] **SULLIVAN Danny.** Portalization and other search trends. In *4th Annual Search Engine Meeting, Boston, 19-20 April 1999.* Available from Internet: <URL: http://www.infonortics.com/searchengines/boston1999/sullivan/sld001.htm>

[21] **VIDMAR Dale, OFTEDAHL Lenora, CHADWICK Terry.** The Invisible Web. In *Librarians in the New Millennium. The Oregon SLA (Special Libraries Association) 2nd Annual Desktop Conference, Oregon, 6-8 February 2001.* [Online]. Available on the Internet: <URL: http://www.sla.org/chapter/cor/desktop/conference_2001/invisible_web/index.html>

[22] **WINDRUM P., SWANN G. M. P.** The `technology fix': developing improved search engines to sustain web growth. In *Proceedings of 1st International Conference on Standardization and Innovation in IT:S12T'99, Aachen, 15-17 Sept. 1999.* Edited by K. Jakobs and R. Williams. NJ: IEEE, 1999. p.133-41.

III-2 Articles de périodiques

[23] 4 Years of Web growth difficult to capture With Accuracy. *Interactive Services Report,* April 17 1998, vol. 19, Issue, 8, (1081 words).

[24] EMMS details four years of attempts to measure Web growth. *Electronic Mail & Messaging Systems,* April 17 1998, vol. 22, Issue 8, (1257 words).

[25] Internet, Ebooks discussed at Online World: The Invisible Web. *Advanced Technology Libraries,* Nov. 2000, vol. 29, no. 11, p. 1, 8.

[26] OCLC Research Project measures scope of the Web. *Advanced Technology Libraries,* Oct. 2000, vol. 28, no. 10, p. 1. Also available from Internet: <URL:

http://www.findarticles.com/cf_0/m3336/9_16/56260081/p1/article.jhtml?term=%22web+statistics%22>

[27] OCLC researchers measures the World Wide Web. *OCLC Newsletter,* November/December 2000, no. 248, p. 25.

[28] OCLC tracks Web. *Information World Review,* November 1999, vol. 152, p.3.

[29] The `deep web': vast and uncharted until now. *eSchool News,* 1 September 2000, vol. 3, no. 9, p. 32.

[30] The Invisible Web. *Advanced Technology Libraries,* November 2000, vol. 29, no. 11, p. 1, 8.

[31] Web contains 7 millions unique sites, says OCLC. *Advanced Technologie Libraries,* Dec. 2000, vol. 29, no. 12, p.1-2.

[32] **ABREU E.** Diving into the deep Web -- Companies are hunting for buried treasure in online databases inaccessible to conventional search engines. *The Industry Standard,* 11 September 2000, vol. 3, no. 35, p. 119. Also available from Internet: <URL: http://www.thestandard.com/article/0,1902,18134,00.html>

[33] **ALBERT Réka, JEONG Hawoong, BARABASI Albert-László.** Diameter of the World-Wide Web. *Nature,* 9 September 1999, vol. 401, no. 6749, p. 130. Also available from Internet: <URL: http://www.nd.edu/~networks/Papers/401130A0.pdf>

[34] **ANDREWS Whit.** Challenge for spiders: searching invisible Web -- Dynamically created pages hard to archive. *WebWeek,* 3 February 1999, vol. 3, no. 3, p.48-51.

[35] **BAR-ILAN Judith.** Data collection methods on the Web for informetric purposes – A review and analysis. *Scientometrics,* 2001, vol. 50, no. 1, p. 7-32.

[36] **BAR-ILAN Judith.** The Web as an information source on informetrics? A content analysis. *Journal of the American Society for Information Science,* 15 March 2000, vol. 51, no. 5, p. 432-443.

[37] **BRADLEY P.** Virtual libraries and Internet searching. *Online & CD-ROM Review,* December 1999, vol. 23, Issue 6, p. 353-355.

[38] **CHOWDHURY G. G.** The Internet and Information Retrieval research: a brief review. *Journal of Documentation.,* March 1999, vol. 55, no. 2, p. 209-225.

[39] **CLARK Don.** Inktomi promises improved searches with Web study. *Wall Street Journal (WSJ)*, 18 Jan. 2000, pB10.

[40] **CLARK, D.** Natural language, relevancy ranking, and common sense. *IEEE Intelligent Systems*, July-Aug 1999, vol.14, no.4, p.17-19.

[41] **COHEN L. B.** Searching the Web: the human element emerges. *Choice,* 2000, vol. 37, Issue Suppl., p.17-31.

[42] **DAHN Micheal.** Counting angels on a pinhead: Critically interpreting web size estimates. *Online*, 2000, vol. 24, no. 1, p. 35-40. Also available from Internet : <URL: http://www.onlineinc.com/onlinemag/OL2000/dahn1.html>

[43] **DAHN Micheal.** Spotlight on the invisible web -- A very large portion of the Web that they do not index. *Online,* July 1 2000, vol. 24 no. 4, p.57-62.

[44] **DAHN Micheal.** Spotlight on the invisible Web. *Online,* July 2000, vol. 24, no. 4, p. 57-62. Also available form Internet: <URL: http://www.findarticles.com/cf_0/m1388/4_24/63568434/p1/article.jhtml?term=%22invisible+web%22>

[45] **EDOLS L.** Uncovering the invisible Web. *Online Currents,* October 2000, vol. 15, Issue 8, p.6-8.

[46] **FELDMAN Susan.** New study of Web search engine coverage published (Statistical data included). *Information Today,* September 1 1999, vol. 16, Issue 8, p. 29. Also available from Internet: <URL: http://www.findarticles.com/cf_0/m3336/8_16/55676542/p1/article.jhtml?term=>

[47] **FLOOD Gary.** The web's dark matter. *Information World Review,* September 2000, no. 161, p. 26. Revue online address: <URL: www.iwr.co.uk>

[48] **FOENIX-RIOU, Béatrice.** 800 millions de pages Web, peu indexées par les moteurs. *Netsources,* Juillet-Août 1999, no. 21, p. 1-4.

[49] **FOENIX-RIOU Béatrice.** Le diamètre du WEB. *Netsources,* Novembre/Décembre 2000, no. 23, p. 9.

[50] **FOENIX-RIOU Béatrice.** Le Web "visible dépasse aujourd'hui un milliard de pages. *Netsources,* Janvier/Février 2000, no. 24, p. 1-2.

[51] **FOENIX-RIOU Béatrice.** Quelques précisions sur le Web visible et invisible. *Bases,* Mai/Juin 2000, no. 26, p. 10-11.

[52] **FOENIX-RIOU Béatrice.** Topologie du Web: la théorie du nœud papillon. *Netsources,* Juin 2000, no. 162, p. 9.

[53] **FOENIX-RIOU, Béatrice.** Web invisible: 550 milliards de documents? *Netsources,* Juillet-Août 2000, no. 27, p. 1-3.

[54] **FOENIX-RIOU, Béatrice.** Web invisible: des ressources incontournables. *Bases,* Juillet-Août 2000, no. 163, p. 8.

[55] **GORDON Michael PATHAK Praveen.** Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing & Management.* March 1999, vol. 35, no. 2, p. 141-180. Revue online address:<URL: http://www.elservier.nl/locate/infroproman>

[56] **GREEN David.** The evolution of Web searching. *Online Information Review: The International Journal of Digital Information Research and Use,* 2000, vol. 24, no. 2, p. 124-137. Also available from Internet: <URL: http://general.rau.ac.za/infosci/information/studyguide/First/unit_5/green.htm>

[57] **HARTMAN K., ACKERMANN E.** The invisible Web. *Proceedings of the Computers in Libraries Conference,* March 2000, vol. 15, p. 105.

[58] **HUBERMAN B. A., ADAMIC L. A.** Growth dynamics of the World-Wide Web. *Nature,* 1999, vol. 401, no. 6749, p. 131. Also available from Internet: <URL: http://www.nd.edu/~networks/Papers/401130A0.pdf>

[59] **INTRONA, L. and NISSENBAUM, H.** Defining the Web: the politics of search engines. *Computer,* Jan. 2000, vol.33, no.1, p.54-62.

[60] **LAVOIE Brian, O'NEILL Edward, McCLAIN Patrick.** OCLC Office of Research examines Web-accessible information to find order in chaos. *OCLC Newsletter,* no. 230. [Online]. Also available from Internet: <URL: http://www.oclc.org/oclc/new/n230/research.htm

[61] **LAWRENCE S., GILES L. C.** Accessibility of information on the web. *Nature*, 8 July 1999, vol. 400, p. 107-109. Also available from Internet: <URL: http://www.liacs.nl/home/fkremer/webt/articles/accessibility-of-information-on-the-web.pdf>

[62] **LAWRENCE S., GILES L.C.** Searching the Web: General and scientific information access. *IEEE Communication Magazine,* January 1999, p.116-122. Also available from Internet: <URL: http://www.comsoc.org/pubs/free/private/1999/jan/pdf/Giles.pdf> (pour la version PDF) et <URL: http://www.neci.nj.nec.com/~lawrence/papers/search-ieee99/> (pour la version HTML)

[63] **LAWRENCE S., GILES L.C.** Searching the World Wide Web. *Science,* 3 April 1998, vol.280, no.5360, p.98-100. Also available from Internet: <URL: http://www.neci.nj.nec.com/~lawrence/science98.html>

[64] **LAWRENE Steve, PENNOK David M., FLAKE, Gary William.** Persistence in Web references in scientific research. *IEEE Computer.* February 2001, p.26-31. Also available from Internet: <URL:

http://www.neci.nec.com/~lawrence/papers/persistence-computer01/persistence-computer01.pdf>

[65] **McCARTHY S. P.** The search is on for ways to navigate invisible Web sites. *Government Computer News,* 22 February 1999, vol. 37, no. 1, (470 words).

[66] **MOLLOY M., LAWRENCE S., GILES, C.L.** Searching the Web, continued: discussion of Searching the World Wide Web by Steve Lawrence and Lee C. Giles. *Science,* 10 July 1998, v. 281, no. 5374, p. 176-177.

[67] **O'LEARY Mick.** Invisible Web discovers hidden treasures: this new service from Intelliseek opens doors for searchers. *Information Today,* January 2000, vol. 17, no. 1, p.16-18. Also available from Internet: <URL: http://www.findarticles.com/cf_0/m3336/1_17/58565059/p1/article.jhtml?term=%22invisible+web%22>

[68] **OHIO STATE UNIV.** The invisible web – Navigating the Web outside traditional search engines. *Reference & User Services Quarterly,* Winter 2000, vol. 40, no. 2, p.131-134.

[69] **OPPENHEIM C., MORRIS A.M McKNIGHT C.** The Evaluation of WWW search engines. *Journal of Documentation,* March 2000, vol. 56, no. 2, p. 190-211.

[70] **PEDLEY P.** The Invisible Web. *The Library Association Record,* Nov. 2000, vol. 102, no. 11, p. 628-633.

[71] **PRICE Gary.** Myths for today, Hopes for tomorrow. *Searcher*, Jan. 2000, vol. 8, No. 1, pp. 113-116. Also available from Internet: <URL: http://www.infotoday.com/searcher/jan00/price.htm>

[72] **ROGERS Micheal, ODER Norman.** Web estimated at seven million sites. *Library Journal (GLJJ)*, 15 Nov. 2000, vol. 125, no. 19, p16-18, p.2.

[73] **SHERMAN Chris.** The future search: Web of search. *Online*. May 1999, vol. 23, no. 3. Also Available from Internet: <URL: http://www.findarticles.com/cf_0/m1388/3_23/54474833/p1/article.jhtml?term=%22invisible+web%22>

[74] **SNOW Bonnie.** The Internet's hidden content and how to find it. *Online,* 2000, vol. 24, no. 3, p. 61-66. Also available from Internet: <URL: http://www.onlineinc.com/onlinemag/OL2000/sherman5.html> et <URL: http://www.findarticles.com/cf_0/m1388/3_24/61640530/p1/article.jhtml?term=%22invisible+web%22>

[75] **SNYDER Herbet, ROSENBAUM Howard.** Can search engines be used as tools for web-link analysis? A critical view. *Journal of Documentation,* September 1999, vol. 55, no. 4, p. 375-384.

[76] **SULLIVAN Danny.** Crawling under the hood. *Online*. May 1999, vol. 23, no. 3. Also available from Internet: <URL: http://www.findarticles.com/cf_0/m1388/3_23/54474830/p1/article.jhtml?term=%22invisible+web%22>

[77] **SWEETLAND James H.** Reviewing the World Wide Web – Theory versus reality. *Library Trends*, Spring 2000, vol. 48, no. 4, p. 748-768p.

[78] **THELWALL Mike.** Web impact factors and search engine coverage. *Journal of Documentation*, March 2000, vol. 56, no. 2, p. 185-189.

[79] **WILSON K.** Searching the Web: a review and preview. *Online Currents*, 3 October 1999, vol. 14, Issue 8, p. 3.

III-3 Journaux Enligne

[81] **BROOKS Terrence A.** Where is meaning when form is gone? Knowledge representation on the Web. *Information Research* [Online]. January 2001, vol. 6, no. 2. Available from Internet : <URL: http://www.shef.ac.uk/~is/publications/infres/6-2/paper93a.html; http://www.shef.ac.uk/~is/publications/infres/6-2/paper93.html>

[82] **HIRTLE Peter B.** Free and Fee: Future Information Discovery and Access. *D-Lib Magazine* [Online]. January 2000, vol. 7, no. 1. Available from Internet : <URL: http://www.dlib.org/dlib/january01/01editorial.html>

[83] **KOEHLER Wallace.** Digital libraries and World Wide Web sites and page persistence. *Information Research* [Online]. June 1999, vol. 4, no. 4. Available from Internet : <URL: http://www.shef.ac.uk/~is/publications/infres/paper60.html>

[84] **SHERMAN Chris.** The future revisited: What's New with Web Search. *Online* [Online]. May 2000, vol. 24, no. 3. Available from Internet : <URL: http://www.onlineinc.com/onlinemag/OL2000/sherman5.html>

[85] **SHERMAN Chris.** The Invisible Web. *Free Pint* [Online]. 8 June 2000, vol. 64**.** Available from Internet <http://www.freepint.co.uk/issues/080600.htm#feature>

[86] **WIGGINS Richard W.** Coping with the Trillion-Page Web. *Library Journal net connect* [Online]. 15 October 2000. Available from Internet: <URL: http://www.libraryjournal.com/trillion.asp> and < http://www.findarticles.com/cf_0/m1299/11_46/67329304/p1/article.jhtml?term=%22web+growth%22>

[87] **WISEMAN Ken.** The Invisible Web: Searching the hidden parts of the Web. *Learning Technology Review*. [Online]. Available from Interenet: <URL: http://a1552.g.akamai.net/7/1552/51/abc2b343b7691d/www.apple.com/educ

ation/LTReview/fall99/pdf/invisibleweb.pdf> (Pour la version PDF) <URL: http://www.apple.com/education/LTReview/fall99/invisibleweb/> (Pour la version HTML)

III-4 Sites et pages Web

[88] Internet growth data. *Berkeley University Library.* [Online]. <http://info.berkeley.edu/how-much-info/internet/rawdata.html>

[89] Internet Indicators. *OECD* [Online]. < http://www.oecd.fr/dsti/sti/it/cm/stats/indicators.htm>

[90] Sizing the Web: domains, sites, pages. *Caslon Analytics Guide to Web Metrics and Statistics.* [Online]. <http://www.caslon.com.au/metricsguide1.htm>

[91] **BARKER Joe.** Seeing the Invisible Web. *University of Berkeley Library.* [Online]. <URL: http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/InvWebPowerpoint/index.htm>, 29 Jan. 2001.

[92] **BERGMAN Michael K.** White Paper: The Deep Web: Surfacing Hidden Value. *BrightPlanet,* July 2000. [Online]. <URL: http://www.completeplanet.com/tutorials/deepweb/contents04.asp (pour la version HTML); http://128.121.227.57/download/deepwebwhitepaper.pdf (pour la version PDF)>

[93] **BOTLUK Diana.** Mining Deeper Into the Invisible Web. *LLRX.com* [Online]. < http://www.llrx.com/features/mining.htm >, 15 November 2000.

[94] **BOYLE Alan.** The Web's bigger than you think: A new statistical survey estimates that the World Wide Web contains at least 320 million pages - far more than previously thought. *ZDNet,* 3 April 1998. [Online]. <URL: http://www.zdnet.com/zdnn/content/msnb/0403/304045.html>

[95] **BRODER Andrei, KUMAR Ravi, MAGHOUL Farzin, …** Graph structure in the Web. Study by Altavista, IBM, Compaq. *IBM corp.* [Online]. < http://www.almaden.ibm.com/cs/k53/www9.final/>

[96] **GOODRUM Abby A., McCAIN Katherine W., LAWRENCE Steve.** Computer science literature and the World Wide Web. *Preprints, 2001.* [Online]. <http://www.neci.nec.com/~lawrence/papers/cs-web01/cs-web01.pdf>

[97] **HOWARD John D.** An analysis of security incidents on the Internet 1989-1995: Chapter 2: Internet Characteristics. *CERT® Coordination Center Research.* [Online].
<http://www.cert.org/research/JHThesis/Chapter2.html>

[98] **Inktomi Corp.** Web surpasses one billion documents: Inktomi and NEC Research Institute complete first Web search. *Inktomi* [Online].
< http://www.inktomi.com/new/press/2000/billion.html>, 18 January 2000.

[99] **Inktomi Corp.** Inktomi Webmap. *Inktomi* [Online].
<http://www.inktomi.com/webmap/>, 18 January 2000.

[100] **Internet Software Consortium.** Distribution of Top-Level Domain Names by Host Count: July 2000. *Internet Software Consortium. [Online].*
<URL: http://www.isc.org/ds/WWW-200007/index.html>, Jul. 2000.

[101] **Internet Software Consortium.** Distribution of Top-Level Domain Names by Host Count: January 2001. *Internet Software Consortium. [Online].*
<URL: http://www.isc.org/ds/WWW-200101/dist-bynum.html>, Jan. 2001.

[102] **Internet Software Consortium.** Internet Domain Survey, July 2000: number of hosts advertised in the DNS. *Internet Software Consortium. [Online].*
<URL: http://www.isc.org/ds/WWW-200007/index.html>, Jul. 2000.

[103] **Internet Software Consortium.** Internet Domain Survey, January 2001: number of hosts advertised in the DNS. *Internet Software Consortium. [Online].*
<URL: http://www.isc.org/ds/WWW-200101/index.html>, Jan. 2001.

[104] **KENNON Julie, JOHNSON Aimee.** Internet Exceeds 2 billion pages. *Cyveillance. [Online].*
<URL: http://www.cyveillance.com/newsroom/press/000710.asp>, 7 July, 2000.

[105] **LAVOIE Brian, O'NEILL Edward, McCLAIN Patrick.** A methodology for sampling the World Wide Web. *OCLC Office of Research.* [Online].
<http://www.oclc.org/oclc/research/publications/review97/oneill/o'neillar980213.htm

[106] **LAVOIE Brian, O'NEILL Edward, McCLAIN Patrick.** Web sites: concepts, issues and definitions. *OCLC Office of Research.* [Online].
<http://wcp.oclc.org/>

[107] **LAVOIE Brian, O'NEILL Edward, McCLAIN Patrick.** Web statistics. *OCLC Office of Research.* [Online].
<http://wcp.oclc.org/>

[108] **LAWRENCE Steve, GILES C. Lee.** Accessibility and distribution of information on the Web. *NEC Research Institute.* [Online].

<http://wwwmetrics.com>

[109] LAWRENCE Steve, GILES C. Lee. How big is the Web? How much of the Web do the search engines index? How up to date are the search engines?: New 1999 study on the Accessibility and distribution of information on the Web. *NEC Research Institut.* [Online].
<http://www.neci.nj.nec.com/homepages/lawrence/websize.html>

[110] MURRAY Brian H., MOORE Alvin. Sizing the Internet: A White Paper. *Cyveillance*, July 10 2000. [Online].
<URL:
http://www.cyveillance.com/us/contact/form_whitepapers.asp?sc=13>

[111] SHERMAN Chris. The Invisible Web. *About: The Human Internet.* [Online].
<http://www.websearch.about.com/internet/websearch/library/weekly/aa061199.htm>, 11 June, 1999.

[112] SMITH Ian. The Invisible Web: Where Search Engines Fear to Go. *Power HomeBiz Guides.* [Online].
< http://www.powerhomebiz.com/vol25/invisible.htm >, cop, 2001.

[113] SULLIVAN Danny. Invisible web gets deeper. From the Search Engine Report. *Search Engine Watch.* [Online].
<URL: http://www.searchenginewatch.com/sereport/00/08-deepweb.html>, Aug. 2, 2000.