Bibliothèque Centre Pompidou
publique d'information

Les consultations “ libres ” d’Internet à la Bpi : Enquête exploratoire

Muriel Amar, Bruno Béguet

2008

Sommaire

Introduction

Voici bientôt dix ans que la Bpi propose à ses lecteurs un accès à Internet : c'est en juin 1995 que les six premiers postes de consultation d'Internet apparaissent dans les espaces de la bibliothèque. Un an après l'ouverture du service, un premier portrait de l'usager-type est dressé par Anne-Sophie Chazaud [CHA 1997] : c'est un homme, jeune et bachelier, généralement issu d'une filière scientifique, assidu de la bibliothèque et utilisateur des collections imprimées, qui déclare utiliser Internet à des fins documentaires, passant sous silence d'autres usages plus ludiques (mais néanmoins visibles) du réseau.
Puis les postes Internet libre déménagent avec une partie des collections et c'est à Brantôme que se met en place le premier système, alors manuel, de réservation des postes (durée de consultation d'une heure). Le public se diversifie, même s'il reste très masculin ; il s'intensifie aussi, notamment grâce aux sessions de formation organisées par la bibliothèque.
La réouverture de la bibliothèque en 2000 se signale par une montée en charge significative du nombre de postes réservés à la consultation d'Internet (50) et par la mise en place progressive d'un système de contraintes techniques et réglementaires destinées à rendre viable la consultation publique d'Internet (système de réservation, charte d'utilisation, interdiction de certains protocoles, etc.).
Parallèlement, les sessions de formation disparues, les bibliothécaires perdent le contact avec les usagers de ce service : très peu de questions, autres que techniques, aux bureaux d'information ; la médiation se réduit à une transaction de tickets de réservation, le plus souvent effectuée sans paroles. Ce service échappe, encore plus que d'autres, à l'appréciation, alors que les seuls chiffres du nombre de réservation indiquent clairement une utilisation très intensive de l'Internet à la Bpi (en moyenne 17986 réservations mensuelles en 2003).

A la faveur d'un contexte rappelé en note méthodologique, la bibliothèque s'est dotée, depuis juin dernier, d'un instrument de recueil des traces de consultations menées sur les postes Internet libre de la bibliothèque.
Qu'indiquent les données recueillies ? Comment les utiliser ? Sous quels angles les analyser ?
Tel est l'objectif de cette étude exploratoire qui n'entend pas épuiser tous les enseignements qu'il est possible de déduire de l'outil à disposition mais qui s'est attachée :

- d'une part à examiner, de façon fine, les données en tâchant de faire émerger les problèmes méthodologiques liés au type de traces recueillies ;
- d'autre part à exprimer les principales tendances des consultations " libres " d'Internet à la Bpi en partant, non pas de grilles de lecture *a priori*, mais de la seule observation des données collectées.

Enfin, des référentiels d'analyse sont proposés en troisième partie de ce document qui permettent de contextualiser, sous différents aspects, les principales tendances précédemment qualifiées.

I- Note méthodologique

I.1.Rappel du contexte

A l'arrivée en stage de Matthieu Renault, élève ingénieur en informatique de l'Université technologique de Compiègne, le service informatique de la Bpi a proposé qu'une première partie de son travail soit consacrée au développement d'une application permettant le recueil des sites consultés à partir des postes Internet libre et Fédération de la bibliothèque.
L'objectif visé était celui de la sécurité : " en cas de soupçon de dérive d'utilisation des postes, il faut pouvoir sortir des statistiques sur une période précise " [CAH 2001, p. 1].

A ce premier objectif s'en est ajouté un second, celui d'identifier les types de consultations menées sur les postes Internet libre. Ce second objectif s'inscrivait dans le cadre d'une recherche, plus large, de nouvelles modalités d'accès aux ressources électroniques à la bibliothèque [GEO 2004].

Cette entreprise trouve également sa place dans un contexte plus général d'interrogation sur les usages des collections et des services de la Bpi, souci qui s'est en particulier traduit par l'organisation de deux " semaines-test " en 2001 et 2003 [SEM 2003].
Un groupe de travail s'est constitué autour du service informatique avec des représentants du DIE et de la cellule Evaluation pour définir un cahier des charges [CAH 2004]. Ce cahier des charges prévoyait notamment que soient possibles des comparaisons entre les sites consultés depuis Internet libre et les sites proposés par la bibliothèque dans ses différents réservoirs : fédération, annuaire de sites et signets de la France contemporaine.

I.2. Bref descriptif de l'application développée [REN 2004]

A l'heure actuelle, l'application développée recueille quotidiennement les " logs " (traces de connexion) transitant par le proxy de la bibliothèque (serveur dédié à l'accès au réseau Internet).
Une série de traitements permet :

- d'identifier le type de poste émetteur : de façon précise pour les postes Internet libre (pourvus de numéro IP fixe), par " défaut " pour les autres postes publics (fonctionnant sur la base de numéros IP " flottants ") ; ce second groupe comprend aussi bien les postes publics Fédération que les postes des bureaux d'information ;

- de transformer le fichier des " logs " initial en un fichier des sites consultés : les logs seuls correspondent aux multiples fichiers nécessaires au chargement d'une page ; ils comprennent aussi des éléments parasites visibles (comme les publicités) et invisibles (comme les compteurs de visites). Un travail de nettoyage et d'agglomération est nécessaire pour reconstituer sous une adresse unique la trace d'un site consulté : nous avons en effet retenu le site et non la page comme unité de référence (troncature des adresses jusqu'au nom de domaine) ; nous avons en outre considéré qu'un site avait été réellement consulté si sa même adresse, reconstituée, était repérable dans un intervalle de 5 secondes.

Les logs nettoyés et transformés en fichier de sites consultés sont ensuite sauvegardés dans une base de données Access. Une interface d'exploitation graphique a été développée avec Excel qui permet, sur la base d'une requête (définition de bornes de date et/ou numéro IP de poste), de disposer de la liste des sites consultés avec le nombre d'occurrences par site, une distinction par type de poste (libre/fédération) et une indication des différences et similitudes entre les consultations " libres " et les sites répertoriés (dans les différentes sélections existant dans la bibliothèque).

I.3. Cadrage de l'étude exploratoire

Cette étude exploratoire a pour objectif d'identifier les types de consultations menées sur les postes Internet libre (partie 2) et de se doter de référentiels de comparaison (partie 3).

A.Univers de référence et premières hypothèses :

S'il existe des études sur les usages domestiques de l'Internet, rares sont celles qui portent sur les usages effectifs en lieux publics (comme les cybercafés), aucune (à notre connaissance) sur ceux pratiqués dans les bibliothèques françaises.
L'application dont nous disposons ne nous permet pas en outre de mener une réelle étude des usages : nous ne disposons que de " traces d'usage ", et non de l'ensemble des pages consultées, encore moins des pratiques développées à partir de ces pages au cours d'une session individuellement notifiée.
L'un de nos objectifs étant de contextualiser les types de consultations, il nous fallait disposer d'univers de référence que nous avons en première instance déterminés en fonction des hypothèses suivantes :

a) les consultations d'Internet à la bibliothèque sont-elles représentatives des consultations du web en général ? Autrement dit, quel est l'impact du lieu " bibliothèque " sur le profil des consultations ? Pour traiter cette question, nous avons recueilli des données de consultation par sites diffusées par la société NielsenNetRating (associée désormais à Médiamétrie) ;
b) les consultations d'Internet libre sont-elles différentes, complémentaires, proches de celles menées sur les postes Fédération ? Autrement dit, quel est l'impact du " libre " sur le profil des consultations ? Pour traiter cette question, nous avons mené un travail de comparaison avec les consultations menées sur les postes Fédération ;
c) les consultations d'Internet libre constituent-elles un pendant " électronique " des consultations des collections offertes à la bibliothèque ? Autrement dit, quel est l'impact du média Internet sur le profil des consultations ? Pour traiter cette question, nous nous sommes intéressés à la forme des " parcours ", comparant ceux suivis dans les collections de la bibliothèque avec ceux suivis sur le web.

B.Choix de la période :

Nous avons retenu l'ensemble des consultations du mois de juin 2004 (24 jours ouvrables), nous situant dans une période de l'année comparable avec celle de la semaine-test 2003 (fin mai-début juin). Ce choix a été par ailleurs motivé par le fait que juin constituait la première période pour laquelle nous disposions de données fiables sur l'intégralité d'un mois.
On notera cependant que le mois retenu est celui qui enregistre le nombre de réservations de postes Internet le plus faible de l'année (jusqu'à présent) :

	Janv.	Fév.	Mars	Avril	Mai	Juin	Juil.	Août	Sept.	Oct.
2003	19773	17470	18930	17482	16588	17330	18264	19107	17613	18692
2004	18969	16932	18104	18525	17806	**16806**	19484	18505	18210	19725

Rappelons enfin, comme le fait observer Françoise Gaudet, que le mois de juin est régulièrement atypique en termes de fréquentation : à une première quinzaine du mois marquée par une forte affluence, notamment estudiantine (dernier coup de collier avant les examens), succède une deuxième quinzaine caractérisée par une fréquentation plus faible, annonce de la trêve estivale comme du renouvellement du profil de lecteurs à la bibliothèque.

C.Choix du corpus :

Nous avons retenu pour analyse uniquement les données pouvant constituer des " traces d'usages ", c'est-à-dire les sites consultés et les données quantitatives de consultations.
On appellera " sites consultés " les sites ramenés à leur nom de domaine suite aux opérations de troncature[1], de regroupement[2] et de nettoyage[3].
On appellera nombre de consultations le nombre d'occurrences enregistrées pour l'ensemble des sites reconstitués (nombre total de requêtes) et/ou pour une URL " normalisée " (occurrences strictes enregistrées pour un site). Dans ce document, les termes " consultations ", " requêtes " et " occurrences " sont donc considérés comme équivalents.

Données du mois de juin

	Internet libre
Total requêtes	870860
Nbre sites différents	58217
Nbre de postes	50

D.Constitution de l'échantillon :

Compte tenu des volumes à analyser, nous avons travaillé par échantillonnage. Deux sous-ensemble ont été constitués :

- un échantillon de travail constitué des mille premiers sites les plus consultés au mois de juin 2004 ;
- un échantillon d'analyse constitué des 400 sites les plus consultés au mois de juin 2004.

Pour obtenir l'échantillon d'analyse fiable de 400 sites, nous avons examiné une à une les 630 premières adresses (dans l'ordre décroissant du nombre de requêtes), éliminé les chargements involontaires, regroupé les adresses qui conduisaient au même site[4].

Notre échantillon d'analyse exclut donc les sites involontairement chargés et comptabilise sous une seule et même adresse les sites qui étaient appelés par des adresses légèrement différentes dans l'ensemble de l'échantillon de travail (1000 premiers sites).

Ces échantillons peuvent être considérés, du point de vue des requêtes, comme représentatifs de l'ensemble des consultations du mois de juin 2004. Ils le sont beaucoup moins du point de vue des sites consultés et rendent mal compte de l'éparpillement et de la diversité des adresses sollicitées :

---

[1]. Reste que, dans certains cas, comme celui des sites personnels (cf. http://perso.club-internet.fr), les URL obtenues ne constituent pas des " traces d'usage " : il faudrait se situer au niveau de la page.

[2] . Un " même " site (une même entité éditorial ou intellectuelle) peut se donner sous des adresses différentes (par exemple, le portail chinois www.sina.com.ch se décline en sous-sites : sport.sina.com.ch, news.sina.com.ch, astro.sina.com.ch, ent.sina.com.ch, etc.) ; on a regroupé les occurrences de chargement de ce site sous une même adresse " normalisée " à sa forme la plus commune (en l'occurrence, www.sina.com.ch). Le programme de Matthieu Renault ne permet pas d'automatiser ces regroupements.

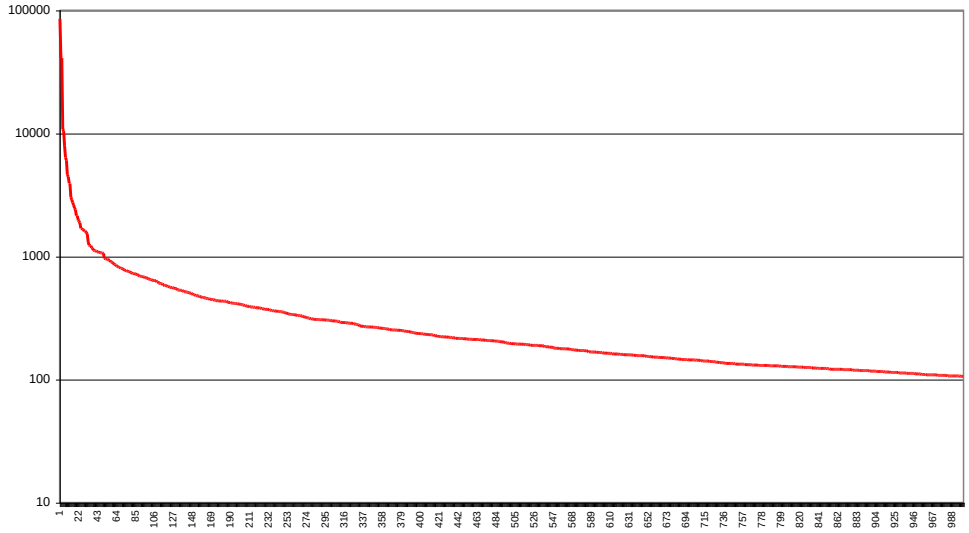
[3] .Toutes les adresses chargées à l'insu de l'internaute ont dû être éliminées manuellement et viendront enrichir la " liste noire " du programme de Matthieu Renault : elles concernent pour l'essentiel des pop-up, bannières, compteur de visites, hébergeurs de sites, etc.

[4] . Ce travail s'est révélé plus long et complexe que prévu, notamment parce que notre base de départ est constituée des URL de pages " tronquées " à leur nom de domaine et non des noms de sites eux-mêmes : nous sommes donc confrontés à une extraordinaire variété dénominative (par exemple, comment savoir que " telefoot.tf1 " et "sports.tfl " sont en fait des adresses différentes pour consulter le même site " eurosports.tf1 " ?) ; nous devons en outre traquer les pages, très nombreuses, de publicités ou de compteurs de visites qui, empruntant souvent des dénominations tout à fait inoffensives, nous obligent toujours à une vérification manuelle.

les 1000 premières adresses totalisent 61% des requêtes, alors qu'elles représentent moins de 2% de toutes les adresses différentes recensées pendant le mois.

	Nombre de requêtes	% / total requêtes	% / total adresses
1000 premières adresses…	528 498	61 % requêtes	1,71 %
…parmi lesquelles 630 premières adresses examinées…	480 485	55,2 % des requêtes	1,08%
…pour constituer un échantillon des 400 premiers sites	418 549	48,1 % requêtes	0,68 %

La répartition des consultations des 1000 premières adresses est la suivante :



Nombre de consultations des mille premiers sites consultés (échelle logarithmique)

Les 50 adresses qui ont connu plus de 1000 requêtes, et qui représentent moins de 0,1 % de toutes les adresses appelées, concentrent à elles seules 34% de toutes les requêtes.

Ce phénomène relève d'une loi dite de puissance (*power-law distribution*) : une faible part de sites concentrent la majorité des requêtes, tandis qu'un nombre important de sites ne sont vus que très rarement [BEA 2004, p. 31].

I.4.Méthode d'analyse

A notre connaissance, il n'existe pas d'étude des consultations libres du web menées en bibliothèque. Nous avons donc constitué une méthode d'analyse *ad hoc* et récursive :

A.Première visite des sites : description libre du contenu ; identification des langues et des

pays producteurs des sites

A ce premier stade, deux difficultés nous sont apparues : l'analyse ayant été menée en août-septembre sur des sites consultés en juin, quelques sites n'étaient plus accessibles (ils ont été extraits de l'échantillon d'analyse) ; beaucoup de sites n'étant ni francophones ni anglophones, nous avons parfois eu du mal à identifier les langues et bien souvent les contenus précis. Nous avons donc fait appel à des informateurs pour la caractérisation des sites en russe et en chinois, ces deux langues étant très représentées dans notre échantillon. Certains sites publiés en plusieurs langues ont été identifiés comme tels (multilingues), sans mention plus fine des langues de publication.
Cette identification précise des langues vise à compléter les informations trop souvent difficiles à interpréter que délivrent les extensions (en particulier : " .com " , " .net " et " .org "), sur lesquelles des statistiques sont produites dans le programme de Matthieu Renault.
La caractérisation des pays s'est révélée plus problématique et souvent guère pertinente : ce critère ne sera d'ailleurs pas utilisé systématiquement et pour lui-même dans l'analyse, mais comme éclairage complémentaire de la répartition par langue.

B.Deuxième visite des sites : caractérisation des sites selon le contenu et l'usage

A partir des descriptifs obtenus en première visite de sites, nous avons établi une typologie d'analyse. A une catégorisation thématique, déterminée par le seul contenu documentaire des sites, nous avons préféré une typologie mixte reposant aussi sur les usages, qui nous paraît mieux rendre compte de la diversité des sollicitations dont l'Internet libre est l'objet à la Bpi.

Neuf grandes catégories, entre lesquelles ont été répartis les 400 sites de l'échantillon, ont ainsi été dégagées. Le tableau ci-dessous détaille les critères sur lesquels cette répartition s'est fondée.

Catégories	Types de sites concernés
Actualités	Médias tous supports (sites de presse papier, radio, télévision, presse en ligne, etc.)
Documentaires	Tous sites à vocation documentaire, quel qu'en soit le contenu, sites institutionnels à vocation documentaire (ex : organisations internationales)
Enseignement	Sites d'institutions d'enseignement (principalement niveau supérieur), informations sur la formation scolaire et universitaire (type Onisep, CIDJ)
Loisirs	Jeux en ligne, humour, horoscope, émission de téléréalité, sites pornographiques…
sites Marchands	Vente en ligne de biens et services (voyages, téléphonie mobile, marchandises culturelles, vente aux enchères…), comparateurs de prix
Portails et moteurs	Tous portails et moteurs de recherche, y compris avec une dimension "actualités" importante, y compris spécialisés
Pratique	Recherche d'emploi, de logement, d'itinéraires, annuaires, petites annonces, sites institutionnels à vocation pratique (ex : CAF,Préfecture Paris)
Rencontres	
Services web	Messageries, chats, blogs - tous prestataires de services Web

Cette catégorisation opératoire a cependant ses limites : les frontières entre certaines catégories sont poreuses et nombre de sites relèvent en fait d'une double affectation. L'incertitude est particulièrement manifeste entre *Portails* et *Actualités*, ce qui nous conduira, dans l'analyse, à détailler la catégorie " Portails et moteurs ". Elle existe également entre *Pratique* et sites *Marchands*, puisque nombre de sites marchands peuvent servir une recherche pratique sans transactions commerciales (horaires de train, recherche bibliographique ou discographique dans les sites de librairies en ligne, etc.). De la même façon, on sait que la plupart des *Portails* proposent un ensemble étendu de services de communication : messagerie, chats, blogs, etc. Plus généralement, dans ce type de site, les trois pratiques majeures permises par l'Internet aujourd'hui – Recherche, Communication et Consultation – sont de plus en plus intriquées : c'est un travail plus précis au niveau des pages consultées qui permettrait, dans ce cas, d'identifier plus finement les usages effectifs.

Notre échantillon ainsi analysé et codé a été copié sous Access pour traitement statistique.

II-Présentation et analyse des résultats

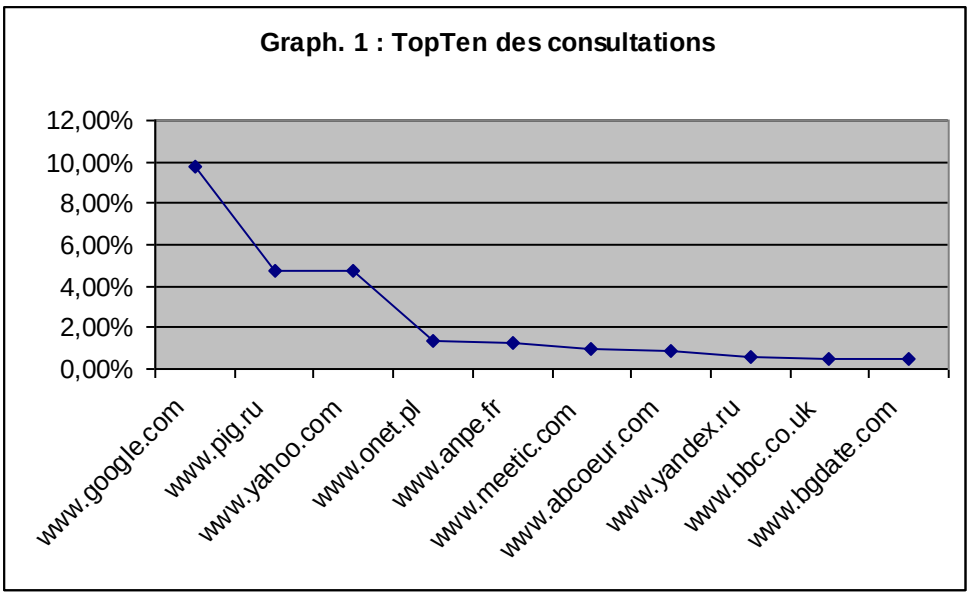
II.1. TopTen des consultations libres du mois de juin

Ordre	URL	Contenu	Langue	Occurrences	%
1	www.google.com	Moteur	Anglais	85547	9.82%
2	www.pig.ru	Jeux en ligne	Russe	41340	4.75%
3	www.yahoo.com	Portail généraliste	Anglais	40958	4.70%
4	www.onet.pl	Portail généraliste et actualités (messagerie, information, etc.)	Polonais	11662	1.34%
5	www.anpe.fr	Pratique (emploi)	Français	10832	1.24%
6	www.meetic.com	Rencontres	Anglais	8377	0.96%
7	www.abcoeur.com	Rencontres	Français	7907	0.91%
8	www.yandex.ru	Portail, messagerie, index web	Russe	5275	0.61%
9	www.bbc.co.uk	Actualités (Presse Audio)	Anglais	4598	0.53%
10	www.bgdate.com	Rencontres	Bulgare	4437	0.51%

Le TopTen est exemplaire :

A-de l'éparpillement des consultations : voir graph. 1 ci-dessous

On constate que, même sur les dix premiers sites les plus consultés, seuls 5 sites représentent plus de 1% des consultations totales. Nous retrouvons là une caractéristique distinctive du web, intimement liée à son volume et à son caractère ouvert : très rares sont les sites qui parviennent à concentrer sur eux un nombre important de consultations [Haering 2002].



B- de la diversité linguistique du web :

Sur les 10 premiers sites, on note 5 langues différentes et une faible représentativité du français (2 sites exclusivement en français). Là encore, cette première tendance, qui sera à confirmer, est représentative du web qui compte environ 3% de sites francophones (source : GlobalReach).
Rappelons, que, pour les moteurs de recherche, et dans une moindre mesure pour les portails, la langue de publication du site n'est pas toujours représentative de la langue réellement utilisée par les internautes : c'est notamment le cas pour Google, site nativement en anglais, mais utilisable en plusieurs langues.

C- de la diversité des usages :

- *Prédominance des moteurs et portails*, attendus comme points de départ pour une navigation libre ; prédominance que l'on retrouve dans les TopTen concernant les consultations des internautes en général (et en France).

NetRating value France: Top 10 Parent Companies Month of August 2004 Home/Work Panel

Property Name	Unique Audience (000)	Reach %	*Time Per Person*
Microsoft	12,929	73.77	02:34:17
Wanadoo	10,308	58.82	00:59:57
Google	9,779	55.80	00:25:28
Iliad - Free	8,260	47.13	00:25:51
Yahoo!	7,335	41.85	00:53:58
PagesJaunes	5,984	34.15	00:16:23
Time Warner	5,660	32.30	02:25:11
Lycos Europe	4,829	27.55	00:32:14
Tiscali	4,567	26.06	00:11:22
PPR	4,205	24.00	00:15:09

- *Gros trafic sur les sites de rencontres et de jeux*, grands classiques des consultations " web " en 2004.

Le site russe " pig.ru " propose une gamme très large de jeux, notamment des jeux d'argent, mais aussi des chats : par nature, ce type de site génère un nombre de requêtes importants. La suite de l'analyse confirmera ou pas la prédominance de consultation pour ce type de sites.
La bibliothèque représente aussi, avec le score obtenu dans nos murs par le site " meetic ", une

tendance très actuelle du boom des sites de rencontre sur Internet, dont, par deux fois en 2004, *Le Monde* s'est fait l'écho : si " 3 millions d'hommes et de femmes en France fréquentent des sites de rencontre " (édition du 15 février 2004), pourquoi pas les lecteurs de la Bpi ?

- *Consultation importante de sites d'actualités* avec le bon score de la BBC, qui ne paraît pas, là encore, spécifique à la Bpi.

NetRating value United Kingdom: Top 10 Parent Companies / Month of August 2004

Property Name	Unique Audience (000)	Reach %	*Time Per Person*
Microsoft	15,527	71.28	02:19:20
Google	10,856	49.84	00:17:29
Yahoo!	9,499	43.61	01:05:43
eBay	7,293	33.48	02:15:39
BBC	6,547	30.06	00:30:19
Time Warner	6,311	28.97	03:01:20
Wanadoo	4,981	22.87	00:21:31
Amazon	4,607	21.15	00:21:29
Ask Jeeves	4,473	20.53	00:11:40
British Telecom	4,039	18.54	00:12:54

- Très bon score de *l'ANPE*, premier site d'information en français consulté depuis les postes Internet libre. La bonne position de l'ANPE est-elle isolée ou au contraire représentative d'un segment de consultation bien spécifique à la bibliothèque ?

Commentaires généraux :

- Sur les dix premiers sites consultés à la Bpi, 4 seulement sont des portails, c'est-à-dire des sites points de départ ; 6 sont des sites " d'arrivée " : il y a donc, en tout cas, sur la base du TopTen, une légère prédominance d'une approche " balisée " d'Internet (l'internaute sait où il va) sur une approche " exploratoire " d'Internet (l'internaute ne sait pas où aller).
- Sur la base de ce TopwTen, on peut identifier des consultations " canoniques " d'Internet : prédominance des moteurs et portails, bonne représentation des sites de jeux et de rencontres. Des segments de consultation semblent être plus spécifiques à la bibliothèque comme le suggère le bon score obtenu par l'ANPE et par les sites qui ne publient leurs contenus ni en anglais ni en français mais dans des langues généralement peu représentées sur le web.

II.2- Diversité linguistique et géographique

A-Données générales

Les interrogations de Google, au nombre de 85547, n'ont pas été prises en compte dans cette analyse linguistique de l'échantillon : la langue d'interrogation est naturellement inconnue et le caractère " français " ou " anglais " du moteur n'est pas pertinent.

Sur les 400 sites analysés, 24 langues différentes ont été identifiées. En nombre de sites différents comme en nombre requêtes, le français arrive en tête (54 % des sites, 39,5 % des requêtes) suivi de l'anglais (12% des sites, 21 % des requêtes) et du russe (7 % des sites, 18 % des requêtes). L'ensemble constitué par les langues d'Europe centrale et orientale (langues slaves, albanais, roumain) représente un peu plus du quart des requêtes pour moins de 13 % des sites différents.

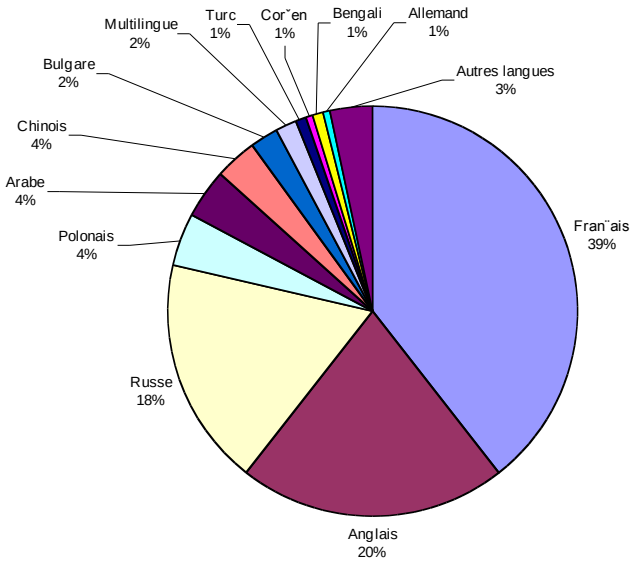
Tableau général des langues, dans l'ordre décroissant des requêtes

Langues	Requêtes	Nbre sites	Requêtes / Sites
Français	131688	215	613
Anglais	70064	48	1460
Russe	60174	29	2075

Polonais	14071	3	4690
Arabe	12398	16	775
Chinois	11659	18	648
Bulgare	6709	7	958
Multilingue	5752	12	479
Turc	2627	7	375
Coréen	2527	5	505
Bengali	2374	3	791
Allemand	2285	5	457
Portugais	1895	2	948
Espagnol	1762	4	441
Néerlandais	1538	5	308
Italien	1232	3	411
Albanais	1201	4	300
Tchèque	797	3	266
Grec	596	2	298
Ukrainien	447	2	224
Roumain	419	2	210
Slovaque	223	1	223
Hindi	209	1	209
Mongol	178	1	178
Persan	177	1	177
Total sf Google	*333002*	*399*	835
Google	*85547*	*1*	*85547*
TOTAL	*418549*	*400*	1046

* Ont été considérés comme " multilingues " les sites proposant sur les mêmes pages 2 langues ou plus (et non les sites proposant des versions en plusieurs langues de tout ou partie du contenu du site).

La répartition par langue des requêtes est la suivante :



B- Focus sur les sites francophones

L'affectation d'une origine géographique à un site n'est pas toujours possible, ni pertinente. Toutefois, il est intéressant de s'interroger sur la part occupée par les sites francophones produits hors de France : sur 215 sites francophones, 35 illustrent de façon incontestable ce cas de figure :

Pays	Requêtes	Sites
France	90177	151
Indéterminé ou non pertinent	21239	29
Algérie	6540	11
Madagascar	5246	7
Côte d'Ivoire	2792	2
Afrique francophone en général	2433	4
Canada	892	2
Maroc	790	2
Sénégal	782	3
Congo	631	3
Belgique	166	1
Total francophonie	131688	215

Ces 35 sites totalisent 15,4 % des requêtes francophones, dont 9 % pour l'Afrique noire francophone et 5,6 % pour le Maghreb.
Sans considération de langue, les sites conçus dans les pays du Maghreb (Algérie essentiellement) représentent un peu moins de 4% de l'échantillon (15786 requêtes pour 21 sites).

Commentaires :

- La forte proportion de consultations de sites non francophones et la variété des langues représentées reflètent la diversité des " communautés linguistiques " qui fréquentent la bibliothèque. Elle doit être rapprochée des observations concernant d'autres services offerts à la bibliothèque, comme les télévisions du monde, la presse étrangère (espace Presse), l'espace Auto-formation.

- Cette diversité linguistique contraste avec la sélection restreinte opérée dans les collections imprimées et électroniques, qui porte, hors le français et l'anglais, sur un petit nombre de langues qui ne sont, de surcroît, pas celles les plus représentées dans l'échantillon : seuls le russe, le chinois et l'arabe, langues très présentes dans les consultations d'Internet libre, sont représentées dans les fonds littéraires.

- Hormis le français et l'anglais, les langues les plus représentées dans les sites consultés à la bibliothèque ne recouvrent pas les langues les plus représentées sur le web en général[5] : il y a donc, de ce point de vue, une singularité des consultations Bpi. La part occupée par les langues slaves, au premier chef le russe et le polonais, illustre cet aspect de façon exemplaire.

II.3.Diversité typologique

A-Données générales

[5]. Selon Global Reach, la distribution par langue de publication sur le web est la suivante (chiffres 2004) : 68,4% de pages en anglais, 5,9% en japonais, 5,8% en allemand, 3,9% en chinois, 3% en français, 2,4% en espagnol, 1,9% en russe, 1,6% en italien, 1,4% en portugais, 1,3% en coréen, 4,6% pour toutes les autres langues. Des données plus précises et complètes seront diffusées en février 2005 par l'Institut des statistiques de l'Unesco qui prépare, dans le cadre de l'Initiative B@bel, un rapport sur l'état du multilinguisme sur Internet : http://portal.unesco.org/.

Le tableau ci-dessous donne, pour chaque catégorie, le nombre de requêtes et de sites concernés, le pourcentage par rapport au total de l'échantillon – ainsi que quelques exemples de sites et le nombre moyen de requêtes par sites selon les catégories.

Catégories	Exemples de sites (par nombre de requêtes décroissantes)	Nbre requêtes	% requêtes	Nbre sites	% sites	Requêtes / Sites
Portails et moteurs	www.google.com - www.yahoo.com - www.onet.pl - www.sina.com.cn - www.privetparis.ru - www.artotal.com - www.ittefaq.com - www.lhotellerie.fr - www.congovision.com	194192	**46,4%**	83	**20,8%**	2340
Actualités	www.bbc.co.uk - www.lagazette-dgi.com - www.aljazeera.net - www.lemonde.fr - www.digitalcongo.net - www.radiofrance.fr - www.nouvelobs.com - www.gazetevatan.com - www.svoboda.org	52392	**12,5%**	86	**21,5%**	609
Loisirs	www.pig.ru - demonscity.combats.ru - www.astrowars.com - www.sex.ru - www.horoscope.fr - anekdot.net - compteur.monjackpot.com	46785	**11,2%**	17	**4,3%**	2752
Pratique	www.anpe.fr - www.caf.fr - www.pagesjaunes.fr - www.pap.fr - www.seloger.com - offres.monster.fr - www.manpower.fr - www.recrut.com - regie.avendrealouer.fr - www.fashion-job.com - www.assedic.fr	38533	**9,2%**	55	**13,8%**	701
Rencontres	www.meetic.com - www.abcoeur.com - www.bdgate.com - www.gaydar.co.uk - www.coucou.org	24005	**5,7%**	15	**3,8%**	1600
Sites marchands	www.voyages-sncf.com - www.ebay.fr - www.avis-verlag.de - www.easyjet.com - kelkoo.fr - www.allworld.ru - www.sonnerie.net - www.sfr.fr - www.fnac.com - www.surcouf.com	20271	**3,8%**	40	**8,8%**	507
Services web	perso.wanadoo.fr - www.canalblog.com - direct.lbe.ru - www.kidon.com - perso.club-internet.fr - www.divan.ru -	15893	**4,8%**	35	**10%**	454
Enseignement	hal.u-paris10.fr - www.univ-paris12.fr - www.cnam.fr -www.univ-paris8.fr - www.france-examen.com - centrale-supelec.scei-concours.org - www.dauphine.fr - www1.admission-prepas.org	13572	**3,2%**	35	**8,8%**	388
Documentaires	www.imdb.com - www.imarabe.org - www.christianity.gr - www.un.org	12906	**3,1%**	35	**8,5%**	380
Total		*418549*		*400*		*1046*

La catégorie " Portails et moteurs " (1/5 des 400 sites) regroupe à elle seule près de la moitié (46,5 %) de toutes les requêtes d'un échantillon constitué des sites les plus sollicités : ce résultat n'est pas surprenant, portails et moteurs représentant pour beaucoup d'internautes un passage obligé. Nous verrons toutefois que des nuances peuvent être apportées à l'intérieur de cette catégorie, qui rassemble des outils de nature diverse, intervenant à un niveau de généralités variable.

80 % des sites et 53,5% des requêtes se répartissent entre les huit autres catégories. La prise en compte de la part occupée par les sites et les requêtes pour chacune d'entre elles, ainsi que la moyenne des requêtes / sites, permet de dégager trois grands types :

A.1.Concentration des requêtes sur un nombre relativement limité de sites : outre la catégorie Portails et moteurs, les catégories Loisirs (11% des requêtes sur 17 sites) et dans une moindre mesure Rencontres (6% des requêtes sur 15 sites) relève de ce régime. Moyenne requêtes / sites : 1000 à 2500.

A.2.Eparpillement des requêtes sur un grand nombre de sites : tel est le cas des catégories Documentaires et Enseignement (au total 6,5 % des requêtes sur 17,5 % des sites), mais aussi Marchands (moins de 4% des requêtes sur près de 9% des sites). Moyenne requêtes / sites : plutôt moins de 500.

A.3.Type intermédiaire : les catégories Actualités et Pratique se situent à la moyenne alliant une grande diversité de sites à une relative concentration des requêtes. Moyenne requêtes / sites : 600 à 1000.

Cette typologie n'est pas propre à la Bpi, il est dans la nature des moteurs et portails, des sites de jeux ou de rencontres de générer des requêtes en nombre. Ce qui doit faire l'objet d'une observation plus attentive est donc :

- l'ordre de classement des catégories, nettement différencié selon la langue (francophones / non-francophones)
- les caractéristiques plus précises des sites rassemblées dans chaque catégorie : disparité des Portails et moteurs, poids des jeux russes dans la catégorie Loisirs, des sites d'offres d'emploi dans la catégorie Pratique, des voyages dans les sites marchands, des universités parisiennes dans la catégorie Enseignement, etc.

B . Catégorisation selon la langue

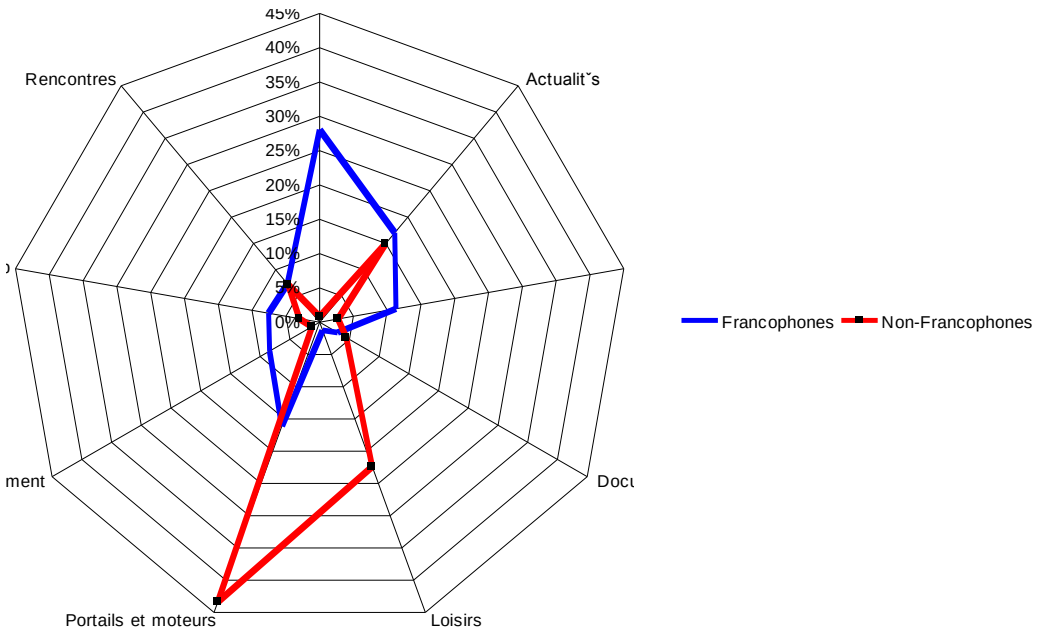
L'analyse qui suit repose sur un échantillon amputé des 85547 requêtes adressées à Google. Le classement par ordre décroissant de sites et de requêtes diffère grandement selon qu'on considère les sites francophones ou les sites non-francophones :

B.1.Sites francophones

Tri décroissant par nombre de sites	**%**	**Tri décroissant par nombre de requêtes**	**%**
Pratique	23,7%	Pratique	**28,1%**
Actualités	**17,2%**	**Actualités**	**17%**
Enseignement	**14,4%**	**Sites marchands**	**16,1%**
Portails et moteurs	**12,6%**	**Enseignement**	**11,3%**
Sites marchands	**12,1%**	**Portails et moteurs**	**8,4%**
Services web	8,4%	Services web	7,5%
Documentaires	5,6%	Documentaires	7,3%
Rencontres	3,3%	Rencontres	3%
Loisirs	2,8%	Loisirs	1,3%

B.2.Sites non-francophones

Tri décroissant par nombre de sites	%	Tri décroissant par nombre de requêtes	%
Portails et moteurs	29,7%	Portails et moteurs	43,4%
Actualités	26,6%	Loisirs	22,4%
Documentaires	12%	Actualités	14,9%
Services web	9,2%	Rencontres	7,1%
Sites marchands	7,6%	Documentaires	4,5%
Loisirs	6%	Services Web	3%
Rencontres	4,3%	Sites marchands	2,7%
Enseignement	2,2%	Enseignement	1,3%
Pratique	2,2%	Pratique	0,8%



Plus de 60% des consultations de sites francophones portent sur les catégories Pratique, Actualités et Portails et moteurs. La catégorie Pratique représente à elle seule le quart des sites, et près de 30% des requêtes. La répartition des sites et des requêtes est homologue: pour chaque catégorie, la part occupée en nombre de sites est proche de celle occupée en nombre de requêtes.

La répartition des sites non-francophones et des requêtes les concernant est tout à fait différente : les requêtes adressées à des Portails et moteurs pèsent pour près de 45% du total, suivies de celles qui portent sur la catégorie Loisirs (près du quart des requêtes pour seulement 6% des sites) et Actualités (15% des requêtes pour près de 27% des sites). La répartition des sites et des requêtes n'est pas homologue : le poids des catégories Loisirs et Rencontres est incomparablement plus important dans les requêtes. La diversité des sites documentaires, des services Web et des sites marchands (entre 7 et 12% des sites pour chaque catégorie) pèse peu en terme de requêtes (entre 2,5 et 4,5% pour chaque catégorie). Enfin, les catégories Enseignement et Pratique occupent une place insignifiante.

C- Focus sur quelques catégories

Focus sur la catégorie Portails et moteurs :
Nous avons déjà évoqué le caractère hybride de certains des sites qui relèvent de cette catégorie. Aux côtés des moteurs et portails généralistes, deux sous-ensembles ont été identifiés : le premier rassemble les sites qui sont tout à la fois des portails et des sites d'actualités, le second concerne les portails spécialisés.

	Sites	Requêtes	Requêtes/Sites
Google	1	85547	85547
Autres Moteurs et portails généralistes	45	39527	878
Moteurs et portails "actualités"	26	65785	2530
Moteurs et portails spécialisés	11	3333	303
TOTAL	83	194192	2340

A l'exception de Google qui a été en moyenne mille fois plus sollicité que les autres moteurs et portails généralistes, ce sont les moteurs et portails " actualités " qui rassemblent le plus grand nombre de requêtes (34% des requêtes de la catégorie, contre 20% pour les portails et moteurs généralistes autres que Google).

La catégorie moteurs et portails d'« actualités » compte en fait un grand nombre de portails de pays étrangers, qui n'ont pour la plupart pas fait l'objet de plus de 500 requêtes dans le mois. C'est leur diversité qui est remarquable : parmi les 400 sites les plus consultés en juin 2004, on dénombre des portails albanais, bulgares, lituaniens, russes, tchèques, chinois, taïwanais, marocains, congolais, ivoiriens, malgaches, brésiliens et paraguayens.
Quant aux moteurs et portails thématiques, ils apparaissent clairement comme répondant à une autre logique qui les apparente de fait à des sites d'autres catégories ; leur ratio requêtes / sites est d'ailleurs proche de celui des sites que nous avons considérés comme " documentaires ". Leur contenu concerne essentiellement les loisirs (cinéma), des renseignements pratiques (astrologie, informations juridiques) ou professionnelles (hôtellerie-restauration). Aucun portail thématique de conception ou de niveau universitaire ne figure dans cette catégorie.

C.1.Focus sur la catégorie Loisirs

Cette catégorie est d'une extrême homogénéité : les jeux comptent pour plus de 93% des requêtes, presque exclusivement concentrée sur le seul site www.pig.ru (41340 sur 43756 reqûetes de jeux). Les sites pornographiques, tous russes, représentent 2% des requêtes de la catégorie – soit 0,23 % de tout l'échantillon, contre plus de 10 % pour les sites de jeux.

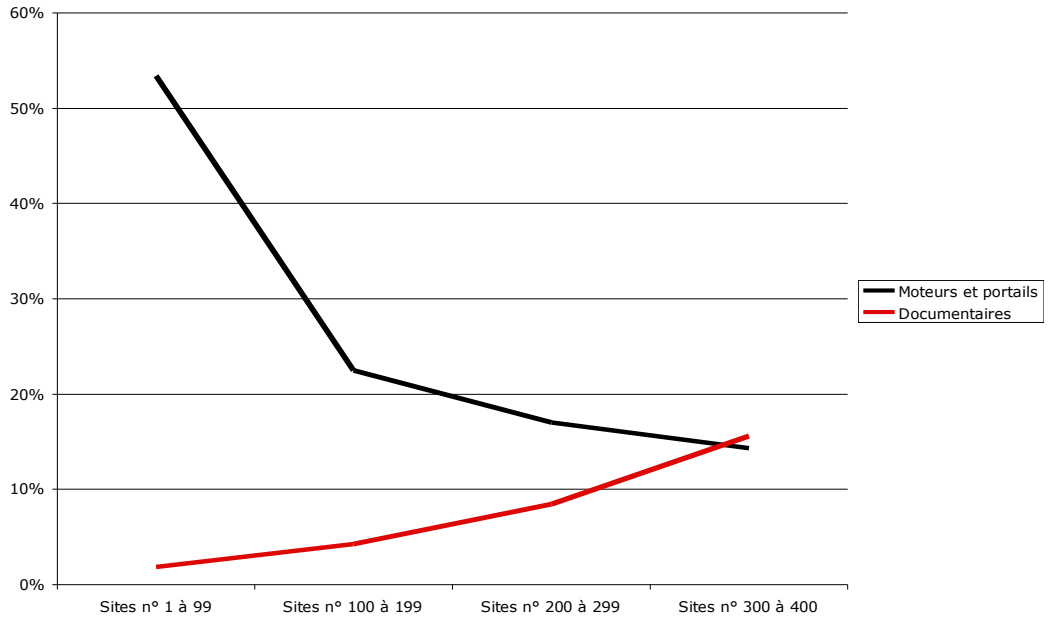
C.2. Focus sur la catégorie Enseignement

Hormis les sites institutionnels incontournables que sont ceux du Ministère de l'Education nationale, du CIDJ, de l'Onisep et du CNAM, l'essentiel des consultations relevant de cette catégorie concernent des établissements parisiens ou des structures parisiennes (53% des requêtes) : toutes les universités parisiennes (sauf Université Paris XI-Orsay) sont représentées, ainsi que les principales grandes écoles de la capitale (Sciences-Po, Centrale-Supelec, Polytechnique, ENS, etc.) et les sites des académies du département (Paris, Créteil, Versailles). Se trouvent, également dans cette catégorie, des sites donnant le résultat de différents types de concours, consultations prévisibles dans la période analysée du mois de juin.
Le profil bien connu du public étudiant de la Bpi se reflète parfaitement dans ce segment de la consultation. Reste une inconnue que fait partiellement apparaître notre échantillon : quelles sont les pratiques développées à partir de ces sites ? La seule consultation ? Ces sites proposent, de plus en plus, des inscriptions, ou pré-inscriptions en ligne, des formulaires ou autres programmes de formation à télécharger, etc. : si les traces de consultation disponibles indiquent que ce type de transaction a été tenté, quelles en ont été les issues ?

C.3.Focus sur les catégories Pratique et Sites marchands

Ces catégories concernent des sites au contenu assez diversifié : pourtant, un petit nombre d'usages concentre une grande partie des requêtes.

Les sites de recherche d'emploi rassemblent 43% de toutes les requêtes de la catégorie Pratique, ce dont ne suffisent pas à rendre compte les 10832 requêtes du site de l'ANPE : 16 autres sites se partagent plus de 6000 requêtes. Les sites de recherche de logement, avec plus de 4000 requêtes , représentent de leur côté plus de 10% des requêtes de la catégorie.
Les sites de voyages rassemblent plus de 36% de toutes les requêtes adressées à des sites marchands. Là encore, les 4015 requêtes du site de la SNCF n'expliquent pas tout : 8 autres sites se partagent plus de 3000 requêtes.
Si la part occupée par la recherche d'emploi confirme un phénomène déjà observé par le biais des



statistiques de consultation de la Fédération, celle des sites de voyages est plus inattendue. Là encore, la question de l'aboutissement des éventuelles transactions, cette fois commerciales, tentées à partir des postes Internet libre se pose.

C.4.Focus sur la catégorie Documentaires

Cette catégorie est par définition, du point de vue des contenus, la plus hétérogène de toute et la plus instable sur la durée. Aucune leçon ne saurait être tirée de la diversité des consultations. Des sites de référence (www.imbd.com pour le cinéma, les sites de l'Union européenne ou de l'Unesco), que l'on peut s'attendre à voir figurer régulièrement parmi les 400 premiers sites consultés, côtoient des consultations conjoncturelles (sur le christianisme en Grèce, la construction des voiliers, la littérature du Penjab, les îles Comores…) qui sont fonction de la fréquentation de la bibliothèque à un moment, voire un jour, donné.
On atteint là les limites de la méthode retenue : faire porter l'étude sur les quelques centaines de sites les plus sollicités conduit à minorer la part occupée dans l'échantillon par les sites documentaires. Une étude centrée sur les consultations documentaires doit reposer sur un échantillon représentatif de l'ensemble des plus de 50 000 sites consultés, et non plus seulement sur les sites concentrant le plus grand nombre de requêtes.
Le graphique ci-dessous illustre la part croissante prise par les requêtes portant sur des sites de la catégorie " Documentaires " : au delà du 300e site, le nombre de requêtes adressé à des sites documentaires l'emporte sur le nombre de requêtes adressé à des moteurs et portails.

Commentaires généraux :

Les tendances, observées à partir du TopTen, se confirment partiellement : la part des moteurs & portails (" recherche d'information " au sens d'orientation sur le web) reste dominante tandis que les pratiques de " consultation " (sites d'actualités et sites pratiques) s'avèrent plus importantes que les pratiques de " communication " (sites de loisirs/jeux, rencontres/chats-services web).
Toutefois, différents profils émergent :

- des profils " francophones " qui privilégient l'accès à une information exploitable hors ligne : vie pratique, enseignement ;
- des profils " non francophones " qui privilégient des services consommables uniquement en ligne : les jeux, le chat, la messagerie, les blogs, etc. Ces profils témoignent de pratiques dans lesquels l'usage du média lui-même est plus recherché que les contenus que ce média peut véhiculer.
- Le cas particulier de l'actualité, contenu à fort taux de renouvellement, mérite d'être isolé : les consultations d'actualités peuvent être assimilées à un service consommable en ligne, engendrant vraisemblablement " des parcours routiniers, rapides et ciblés qui s'apparentent aux modes de consommation des médias traditionnels, comme TV, radio, presse écrite ",

[BEA 2004].

On retrouve, derrière ces différents profils, les grandes catégories d'usagers identifiés dans les enquêtes de fréquentation aussi bien que par l'observation spontanée : la forte présence étudiante se traduit dans la consultation des sites des universités parisiennes, la variété des communautés étrangères, d'origine ou de passage, dans celle des portails et sites d'actualités du monde entier, et tout particulièrement slaves, africains, arabes ou asiatiques. La recherche d'emploi est aussi active sur l'Internet libre que sur la Fédération. Le recours à l'Internet libre pour une utilisation de type "services" est très importante, et transcende nos catégories : les requêtes adressées aux sites que nous avons classés dans les catégories Pratiques et Marchands additionnées à la messagerie et à une bonne partie des requêtes des sites des universités parisiennes, relève de cette configuration. Enfin, s'agissant des utilisations que l'on peut qualifier de " ludiques ", il apparaît clairement qu'elles se réduisent aux jeux, avec une ampleur et une systématicité (quelques sites) qu'on peut d'ailleurs légitimement trouver préoccupante, alors que la pornographie est négligeable.

III.Mise en perspective des résultats : quelques éléments de comparaison

III.1.Comparaison consultations-Bpi et consultations-internautes en général

Est ici brièvement synthétisé un ensemble de remarques dont l'objectif est de contextualiser les consultations observées à la Bpi : sont-elles représentatives des usages du web en général ? révèlent-elles des spécificités sur nos publics ?

Comparées à la répartition linguistique du web (voir note 5), les consultations menées à la Bpi font apparaître de fortes singularités : si les sites en langue russe représentent 1,9% du web, ils représentent 18% des consultations-Bpi. De la même façon, rappelons que le polonais, qui arrive en quatrième place des langues de consultation d'Internet à la Bpi, concentre sur 3 sites 3% des consultations de l'échantillon.

En termes de thématiques, les grandes tendances, relevées par exemple dans Library Trends [SPI 2003] indiquent qu'entre 1997 et 2002, les sujets de recherche ont fortement évolué vers une baisse de fréquentation des sites de loisirs et vers une hausse de consultation des sites de commerce, de voyage et d'emploi notamment.

Si cette tendance est confirmée pour les sites francophones avec une prédominance des sites relevant de la catégorie " pratique " (28,1% des consultations) et de la catégorie " services marchands " (16,1% des consultations), elle n'est en revanche pas avérée pour les sites non francophones, pour lesquels la partie " loisirs " représente 22,4% des consultations contre 1,3% dans le cas des sites francophones.

La concentration des requêtes sur un nombre restreint de sites, manifeste dans le cas des catégories Loisirs (11% des requêtes concernent 4% des sites) et Rencontres (6% des requêtes portent sur 3% des sites), indique des pratiques routinières et bien balisées du web : il semble qu'une partie des internautes Bpi consulte régulièrement un ou plusieurs site(s) déjà connu(s), sans nécessairement chercher à élargir leurs horizons. Ce type de comportement est régulièrement observé dans les études portant sur les comportements d'internautes[6] Dans ces cas-là, la connaissance des sites régulièrement consultés ne passe pas nécessairement par les outils de recherche du web (moteurs ou portails) mais plutôt par les autres médias et, le plus souvent, par les relations informelles développées au sein des communautés d'appartenance des internautes [OUT 2003, p. 176 et suiv.].

En revanche, et essentiellement pour des raisons techniques et réglementaires, les pratiques intensives de communication sur Internet (messagerie, chat, blogs, etc.) restent encore modestes à la Bpi (10% des requêtes) comparées à celles relevées dans d'autres études menées dans le cadre d'usages domestiques ou professionnels[7] : il pourra être sur ce point intéressant de noter la

[6] . [BEA 2004] : " Au sein d'espaces Web a priori non bornés, les internautes dessinent des zones familières de taille restreinte autour de thématiques propres à chacun ", id. [OUT 2003].

[7] [SEN 2002] : si 76% des sessions Internet comprennent nécessairement un passage par le web, 70% des sessions comprennent l'utilisation soit de la messagerie, soit du chat, soit d'un forum.

progression du segment "communication" dans les consultations Bpi dès lors que le protocole de la messagerie y sera autorisé.

III.2.Comparaison des consultations “ libres ” et des consultations “ fédération ”

A-Retour sur la semaine-test 2003

Les analyses conduites par Paule Ruiz [SEM 2003] sur l’offre Fédération ont montré une prédominance des consultations de sites de presse (46,74% des consultations, avec une répartition de 21% pour la presse française et de 24% pour la presse étrangère), suivies, dans une moindre mesure, par les consultations d’usuels (16,06%) et de sites d’emploi (15,30%).
Ces prédominances recouvrent une partie des consultations Internet libre, focalisées, pour les sites francophones sur les sites d’actualités et d’informations pratiques (surtout sur l’emploi).

B-Données de juin 2004 : comparaison des TopTen libres et fédération

TopTen des consultations à partir des postes fédération (bruit retiré)

Top Sites		**Occur**	**%**
www.anpe.fr	Pratique / fre	68342	8,74
www.bpe.europresse.com	Presse / fre	25808	3,30
www.google.com	Moteur / fre ?	22866	2,92
www.lemonde.fr	Presse / fre	20229	2,59
www.lematin-dz.net	Presse / fre Algérie	18909	2,42
e.pagesjaunes.fr	Pratique / fre	16618	2,12
www.nouvelobs.com	Presse / fre	15203	1,94
charlas.elmundo.es	Chat / espagnol	13669	1,75
www.yahoo.com	Portail / fre ?	13554	1,73
www.lequipe.fr	Presse/fre	8804	1.13
Corail.sudoc.abes.fr	Catalogue/fre	6940	0.89

Rappel : Topten Internet libre

Top Sites	CONTENU	Langue	Occur	%
www.google.com	Moteur agglomeré	Anglais / Français	85547	9,82 %
www.pig.ru	Jeux en ligne	Russe	41340	4,75 %
www.yahoo.com	Portail généraliste	Anglais	40958	4,70 %
www.onet.pl	Portail généraliste et actualités (messagerie, information, etc.) : agglomeré	Polonais	11662	1,34 %
www.anpe.fr	Pratique (emploi)	Français	10832	1,24 %
www.meetic.com	Rencontres - agglomeré	Anglais	8377	0,96 %
www.abcoeur.com	Rencontres	Français	7907	0,91 %
www.yandex.ru	Portail, messagerie, index web, agglomeré	Russe	5275	0,61 %

www.bbc.co.uk	Actualités (Presse Audio)	Anglais	4598	0,53 %
www.banex.search.bg	Rencontres	Bulgare	4437	0,51 %

Premiers éléments de comparaison apparaissent :

- L'éparpillement des consultations est moindre dans le cas de la fédération (le volume initialement autorisé est beaucoup plus faible aussi) ;
- La langue dominante est évidemment le français ;
- Le type très largement dominant est la presse, type qui se dégage aussi nettement des consultations " libres ", dès lors que sont exclus les moteurs et portails ;
- L'ANPE reste le premier site consulté dans les murs de la bibliothèques, que ce soit depuis les postes Fédération ou depuis les postes Internet libre (dans ce dernier cas, l'ANPE est le premier site d'information consulté, les portails et les jeux en ligne générant un trafic important du fait de leur nature même).
- Des " détournements " semblent possibles à partir des postes Fédération (à moins que ce ne soit le fait des postes des bureaux d'information) : cf. Google, Yahoo, partie " chat " d'El Mundo.

Si, du point de vue de la volumétrie, les consultations Internet libre et Fédération se recouvrent très peu (environ 5% des requêtes " libres " auraient pu se faire aussi sur les postes fédération), une proximité de profils se dégage : poids de la presse, et notamment poids de la presse étrangère. C'est ici qu'une catégorisation par " type " de sites, au-delà d'une comparaison site à site, se révèle particulièrement intéressante.

C- Sites consultés et sites sélectionnés hors échantillon

En examinant de plus près les consultations Internet libre qui portent sur des sites sélectionnés dans la fédération, on obtient sur les 100 premiers sites " communs ", la typologie suivante :

- Sites de presse & média (tout type) : 26
- Sites pratiques (surtout emploi) : 25
- Sites institutionnels (tout domaine) : 21
- Sites d'enseignement : 14
- On trouve aussi, dans ces consultations, des sites de " bibliothèques " (bnf, sudoc), de collections numérisées (Gallica), des sites " documentaires " sur un sujet ciblé.

Les sites librement consultés qui rencontrent les sélections de la bibliothèque sont de trois types : actualité, pratique et institutionnels.
La part représentée par les sites d'enseignement pourrait être plus importante, si l'offre de la fédération pouvait accueillir les portails des universités parisiennes, et pas uniquement les pages donnant accès aux ressources documentaires des bibliothèques universitaires.

Les autres catégories de sites, notamment les documentaires sur une thématique précise, correspondant à beaucoup trop de sites sur le web pour espérer qu'une coïncidence puisse se produire. Une étude plus fine, sur une durée plus importante et selon d'autres critères d'échantillonnage (cf supra), serait alors nécessaire pour étudier de près les thématiques de recherche développées lors des consultations libres.

D-Commentaires généraux

- Les profils " libres " et les profils " fédération " se recouvrent partiellement sur les segments de fortes consultations que sont la presse et l'information pratique, surtout orientée " emploi ".
- Sont caractéristiques de la consultation libre les pratiques communautaires autour des portails d'actualités étrangères par exemple et les pratiques de communication comme les jeux, les chats, les rencontres.

III.3.Parcours et portraits

A- Méthodologie et objectifs

L'application développée par Matthieu Renault permet de disposer de l'ensemble des sites consultés sur un poste donné et dans un intervalle de temps situé. Seules les sessions à l'ouverture de la bibliothèque sont fiables : par la suite et sur l'ensemble de la journée, on ne peut être sûr, qu'à chaque intervalle de 45 minutes, un nouvel internaute entame un parcours.
Pour des raisons techniques, nous n'avons pu exploiter les données par poste sur le mois de juin 2004 : c'est sur la journée du 7 avril que dix portraits de " session " ont été analysés, voir annexe 1.
Ces données doivent nous permettre de reconstituer des parcours de consultation que nous nous proposons de rapprocher, à des fins indicatives, avec l'étude des parcours parmi les collections de la bibliothèque tels que Barbier-Bouvet a pu les établir [BAR 1986].

B- Commentaires généraux

- Dans l'ensemble, les parcours semblent très orientés, autour d'un seul parfois deux pôle(s) d'intérêt : il semble que le " surf ", la pratique purement exploratoire du web, soit peu développée à la bibliothèque. On remarque que beaucoup de session (8 sur 10) débutent par un accès direct à un site particulier : le recours aux moteurs ou annuaires intervient davantage en cours de session. Les internautes Bpi savent visiblement ce qu'ils cherchent.
  Il est vraisemblable que la durée limitée des sessions encourage les internautes Bpi à " rentabiliser " leur minutes de requêtes. Cependant, de récentes études [SEN 2002 ; OUT 2003] mettent également en exergue cette tendance à développer une pratique balisée de l'Internet, en notant que la pratique purement exploratoire du web fléchit (est-ce un effet d'acculturation au web ?) ou n'est, dans le fonds, qu'une vue de l'esprit …[8]
- Parallèlement, au sein de l'ensemble des sessions généralement " orientées ", on note toujours des digressions ou incursions dans d'autres domaines : une session dominée par la consultation (actualité par exemple) peut également comprendre des pratiques de recherche d'information ou de communication. Il est fréquent que l'internaute passe d'une pratique à l'autre alors même que son itinéraire reste borné.
- Ces quelques remarques, issues d'un panel extrêmement réduit, font écho aux pratiques de lecteurs et aux parcours identifiés par Barbier-Bouvet au sein des collections de la bibliothèque : les lecteurs, dans la bibliothèque comme sur le web, se créent des " univers de familiarité " qui ne les empêchent pas pour autant d'utiliser largement la diversité des services offerts par la bibliothèque (les collections encyclopédiques, l'actualité, le multimédia).
- On remarquera enfin que les langues ne semblent pas constituer un obstacle lors des sessions de consultation qui sont rarement monolingues.

IV-Conclusions

Les principales tendances des consultations "libres" d'Internet que nous avons dégagées à partir d'un échantillon des relevés de juin 2004 sont-elles représentatives des pratiques régulièrement menées sur les postes de la bibliothèque ?
Le TopTen des sites les plus consultés au mois de septembre 2004 le laisserait supposer :

[8] .[OUT 2003, p. 57] : " On a relevé depuis une dizaine d'années déjà les horizons étroits de ces univers personnels de navigation alors que les réseaux, et le web en particulier, nous sont encore présentés comme des bibliothèques infinies où circule une masse d'informations prêtes à être saisies ".

URL	Contenu	Occurrences
www.google.com	Moteur	109555
www.yahoo.com	Portail généraliste	44054
www.pig.ru	Jeux en ligne	27584
top.list.ru	Bruit (lié à un serveur de mails, russe)	18219
www.anpe.fr	Pratique (emploi)	12452
Pokec.atlas.com	Portail en anglais	12386
134.99.100.86	Bruit	9026
www.abcoeur.com	Rencontres	8145
Banex.search	Bruit	6264
www.revefrance.colm	Portail chinois	5868

Fichier original, sans opération de nettoyage et regroupement

Hormis les scores obtenus par Google, Yahoo et dans une moindre mesure Abcoeur (qui détrône désormais Meetic, vraisemblablement en raison de sa gratuité), très représentatifs des usages du web en général, on ne peut être que frappé par l'extraordinaire constance de la fréquentation du site de jeu russe pig.ru et du site de l'ANPE, très clairement spécifiques à la Bpi.
Si le score de l'ANPE peut être directement lié à la politique volontariste de la bibliothèque dans ce domaine, celui du site pig.ru se laisse plus difficilement apprécier : il ne peut être le fait que d'usagers extrêmement réguliers et assidus qui consacrent l'essentiel de leurs sessions (voire plusieurs sessions par jour) sur cet unique site ; or ce site propose notamment des jeux d'argent, dont la légalité sur l'Internet n'est pas totalement avérée : ne devrait-on pas approfondir nos observations sur cet aspect des usages du web et déterminer en conséquence un positionnement de la bibliothèque ? Rappelons que ce site concentre à lui seul, et semble-t-il avec constance, entre 2 et 5 % de toutes les requêtes de l'Internet libre.

Au-delà de ces constantes propres au TopTen, nos analyses de détail ont permis d'affiner les profils de consultations d'Internet à la Bpi, laissant apparaître de fort segments de consultations sur les sites d'actualité, notamment étrangère, et sur les sites francophones d'informations pratiques, tournées essentiellement vers la recherche d'emploi.

Ces deux thèmes, déjà bien couverts par les sélections réalisées dans la fédération, mériteraient vraisemblablement d'être approfondis[9]. Sur ce point, il nous paraît essentiel que l'application informatique de recueil des données puisse servir d'outil de veille aux responsables des domaines qui pourraient trouver dans l'analyse de ces consultations " spontanées " matière à entretenir et actualiser l'offre sélectionnée dans la Fédération.
Pour cela, une amélioration significative des résultats produits est indispensable[10] : le temps passé au retraitement des données dans le cadre de cette étude exploratoire nous a suffisamment montré combien les données issues du web étaient complexes, trompeuses, voire un brin retorses.

En revanche, les limites de notre étude menée sur un panel restreint à la fois dans le temps et dans les volumes sont évidentes dès lors que l'on souhaite rendre compte des thématiques de recherche menées lors des consultations "libres" d'Internet.

[9] . La régularité des sites consultés dans ces deux domaines, et notamment dans celui de l'actualité, révèle là encore vraisemblablement d'itinéraires routiniers cf. [BEA 2004 p. 260] : "il apparaît que lors de ces balayages récurrents de terrains connus, les internautes se connectent à des flux, flux d'information ou flux de communication. Dans les deux cas, le mode de réception n'est pas si éloigné de celui des médias traditionnels : une fois que l'internaute a identifié les canaux qui l'intéressent, c'est par ce biais qu'il consomme du contenu."
[10] . Alimentation (manuelle ? automatique ?) des listes "noires" utilisées dans l'application de Matthieu Renault.

Dans notre échantillon, la catégorie des sites documentaires ne représente que 3% de l'ensemble des requêtes et 8,5% de l'ensemble des sites consultés. Il est évident que ces proportions, notamment en nombre de sites, seraient autres si nous avions exploré l'intégralité des consultations du mois de juin. Il est évident aussi que les consultations de sites liés à des recherches documentaires restent le plus souvent ponctuelles, avec un taux de revisite très faible[11].
Cependant, cette étude mérite d'être poursuivie pour dégager ne serait-ce que les thématiques de recherche dominantes qui motivent les internautes Bpi : un travail sur un recueil annuel des données est alors nécessaire qui procèderait par sondage à un pas à déterminer par mesure statistique en fonction des données globales.

Muriel Amar, Cellule Evaluation
Bruno Béguet, Service des documents imprimés et électroniques
Novembre 2004

[11] . [BEA 2004, p. 247] : les sites qui ne sont vus que dans une seule session représentent les trois quarts des sites vus en moyenne.

Annexe 1 : dix " portraits " de sessions (données d'avril 2004)

Pour chacun des dix postes retenus, on a comptabilisé le nombre de sites consultés sur l'ensemble de la session, on a identifié les modalités d'entrée dans la session : accès direct à un site ou passage par un moteur ou un portail (dit accès indirect), on a ensuite relevé l'ordre d'appel des sites en indiquant la langue utilisée.

Poste 10.3.4.30

- Nombre de sites différents : 4
- Accès direct
- Site de l'association de la communauté de Darfour en France (en arabe)

Puis portail soudanais (annuaire de sites en anglais)
Puis site d'achats de domaine ?? (ou bruit)
Puis site de rechargement de cartes téléphoniques (en français)
Retour sur le site de l'association de la communauté de Darfour
Longue session sur le site d'un mouvement d'égalité et de justice au Soudan (multilingue : arabe, français, anglais, allemand)
⇒ Consultation d'actualité à chaud ; plusieurs langues ; intrusion dans un site marchand ; itinéraire bien balisé.

Poste 10.3.4.32

- Nombre de sites différents : 5
- Accès indirect via moteur (Google)
- Puis accès cidj , anpe (longue session), retour au cidj, passage par site officiel sur l'Europe, retour sur Google, visite d'un site d'entreprise (Lierac beauté, entreprise liée à une offre d'emploi ?), visite d'un site associatif sur l'islam en France

⇒ Session en français orientée pratique/emploi (recherche d'information) avec incursion de consultation d'un site visiblement déjà connu.

Poste 10.3.4.54

- Nombre de sites différents : 2
- Accès direct
- Site d'actualité russe (webzine) : très longue session et portail russe (peut-être chargé en même temps)

⇒ Session en russe mono-site, consultation habituelle d'actualité. Session très balisée.

Poste 10.3.4.39

- Nombre de sites différents : 10
- Accès indirect via annuaire (Yahoo)
- Puis accès site de presse algérienne en arabe, retour sur Yahoo, puis portail d'actualité algérien en français, retour sur Yahoo, puis autre site de presse algérienne en arabe, retour sur Yahoo, puis autre site de presse algérienne en arabe, retour sur Yahoo, autre site de presse algérienne en arabe, retour sur Yahoo, site de presse algérienne en français, retour sur Yahoo, site de presse algérienne en français, retour sur Yahoo, site de presse algérienne en français, retour sur le portail d'actualité algérienne en français, retour sur Yahoo, fin sur le portail.

⇒ Session bilingue français-arabe ; démarche méthodique sur les journaux algériens : à la fois recherche d'information et consultation. Itinéraire de découverte avec un objectif de recherche initial

très précis.

Poste 10.3.4.33

- Nombre de sites différents : 2
- Accès direct
- Deux sites d'actualité russe en russe.

⇒ Session en russe, très balisée, habituelle ? Consultation uniquement.

Poste 10.3.4.31

- Nombre de sites différents : 7
- Accès direct
- Accès sur un site d'entreprise en français (constructeur automobile), puis longue session sur un portail en français et en polonais, puis site en français sur la musique polonaise (avec clip vidéos) puis retour au portail, accès à un moteur de recherche en français, accès à un site de rencontre en français, retour au portail, puis fin de session sur un site de radio polonaise en polonais (avec fichiers audio).

⇒ Session bilingue français-polonais mêlant différentes activités de recherche d'information, consultation, communication. Itinéraire balisé.

Poste 10.3.4.76

- Nombre de sites différents : 5
- Accès direct
- Anpe, puis moteur de recherche (Google), retour sur anpe, puis sur google, puis jobpilot, puis site marchand (petit matériel d'informatique), retour sur jobpilot (très longue session), puis site d'entreprise (consultant), retour sur jobpilot.

⇒ Session monolingue français, centrée sur la recherche d'emploi, brève incursion sur un site marchand. Itinéraire balisé.

Poste 10.3.4.70

- Nombre de sites différents : 16
- Accès direct
- Site de chanteuse en français (princesse Robert), puis site de fan en français (Madonna), chat et forum sur Princesse Robert en français, site officiel d'Indochine en français, retour site princesse Robert en français, site officiel de Princesse Robert en français, portail en japonais, site perso sur Mylène Farmer en français, retour sur le portail japonais, site du magazine sur M. Farmer, site M. Farmer en français, autre site sur M. Farmer en français, encore autre site sur M. Farmer en français, retour sur le site M. Farmer en français, serveur de blog en français (un blog sur Robert), site en anglais sur les arts martiaux, retour sur le premier site consulté, passage par google, retour sur le premier site consulté, site de fan Indochine, passage par google, site de fan Indochine, site de téléchargement d'icônes.

⇒ Session trilingue français, anglais, japonais. Itinéraire très balisé (découverte de nouveaux sites toujours sur le même thème). Dominante consultation et communication.

Poste 10.3.4.75

- Nombre de sites différents : 10
- Accès direct
- Sites d'université (Dauphine, Paris6), passage par Google, sites d'université (Paris 6/7/PMC), retour par google, site d'université, google, anpe, google, anpe, cidj, anpe, google, portail

d'informations sportives en espagnol, google, fin de session sur le portail d'informations sportives en espagnol.

⇒ Session bilingue français et espagnol. Mêle session pratique/enseignement et consultation informations. Itinéraire très balisé.

Poste 10.3.4.77

- Nombre de sites différents : 6
- Accès direct
- Portail d'informations sportives en portugais, autre portail d'informations sportives en portugais, encore autre portail d'informations sportives en portugais ; site en français Assedic (avec traces de transactions), accès à l'application Ravel en fin de session (formulation de vœux d'affectation dans le supérieur).

⇒ Session bilingue français /portugais alternant consultation d'informations ciblées, recherche d'information/vie pratique et usages du web comme mode de communication.

Bibliographie

Documents internes Bpi

CAH 2002. Service informatique. Cahier des charges Statistiques Sécurité Internet. Septembre 2002. 2p.
CAH 2004. Cahier des charges : statistiques sur les accès web à la Bpi. Février 2004. 5p.
GEO 2004. Renouveller l'accès public au web en bibliothèque. Septembre 2004. 13p.
SEM 2003. Béguet Bruno, Dartois Claire, Ruiz Paule. Semaine-test 2003 : rapports 1 et 2. 2003. 25p. et 43p.
REN 2004. Renault Matthieu. Applications statistiques Internet et Fédération. Juin 2004. 25p.

Articles et études

BAU 1996. Baude Dominique. *Internet à la Bibliothèque publique d'information*. Bulletin des bibliothèques de France. 1996. T. 41, n° 1. P. 56-61
BAR 1986. Barbier-Bouvet Jean-François. *L'embarras du choix : sociologie du libre accès en bibliothèque*. Bulletin des bibliothèques de France.1986. T. 31, n°4. P. 294-298
BEA 2004. Beauvisage Thomas. Sémantique des parcours des utilisateurs sur le web. Thèse de doctorat en linguistique. Université Paris X Nanterre et laboratoire "Usages, Créativité, Ergonomie" de France Télécom R&D. 2004
CHA 1997. Chazaud Anne-Sophie. *Usages d'Internet à la Bibliothèque publique d'information*. Bulletin des bibliothèques de France. 1997. T. 42, n° 3. P. 34-40
GAUD 1999. Gaudet Françoise et Evans Christophe. *La Bibliothèque Publique d'Information-Brantôme : un cas de restructuration des publics par l'offre ?* Bulletin des bibliothèques de France. t. 44, n° 4. P. 31-38
HAE 2002. Haering Hélène. *Internet: nouveau média, nouvelles mesures ?* Dossiers de l'audiovisuel. 2002. N° 103. P. 36-38
OUT 2003. Franck Ghitalla, Dominique Boullier, Pergia Gkouskou-Giannakou, et al. L'Outre-lecture : manipuler, (s')approprier, interpréter le web. Paris : Bibliothèque publique d'information - Centre Pompidou, 2003. (Études et recherche)
SPI 2003. Spink Amanda. *Web search : emerging patterns*. Library Trends. Fall 2003. T. 52, n°2. P. 299-306
SEN 2002. France Telecom R&D, NetValue, Université Paris 3, laboratoire LIMSI / CNRS. Projet SensNet : catégorisation sémantique des usages et des parcours sur le Web. http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/sensnet.htm

Sites web

GlobalReach : www.global-reach.biz/globstats/index.php3
Nielsen NetRatings : www.nielsen-netratings.com