Accès au contenu

Daniel Pimienta
Association Réseaux et Développement (FUNREDES)
www.funredes.org

Le world wide web est de fait le plus grand domaine public d'information. Mais à mesure qu'il poursuit la croissance exponentielle de son contenu, les moteurs de recherche perdent leur capacité à en indexer une grande partie ; et la publicité étend sa portée et son influence sur le comportement des utilisateurs de façon de plus en plus directe et perverse lorsqu'ils accèdent au contenu. Parallèlement, des questions comme le contenu ouvert et le droit d'accès à ce domaine public de savoir demeurent importantes, et on constate un certain progrès à cet égard. La démographie du net a fini par évoluer vers une plus grande diversité culturelle et linguistique, annonçant la fin d'une phase initiale et transitoire de domination anglophone, conséquence de son histoire.

Ce chapitre met en garde contre le risque de distorsion des services de recherche en ligne : d'une part, une distorsion de nature culturelle et liée indirectement à une réduction de la couverture ou de la capacité, et d'autre part, une distorsion qui découle directement de l'influence des publicitaires, qui influent délibérément sur les résultats de la recherche.

L'importance du problème est évidente quand on sait qu'un des objectifs légitimes est de permettre l'accès au contenu en ligne dans la langue de chacun et que la fracture numérique est beaucoup plus profonde sur le plan du contenu que sur le plan de l'accès à la technologie. Et les nouveaux internautes qui se connectent sont de moins en moins nombreux à produire du contenu, ce qui montre bien l'importance de la compétence numérique dans la lutte contre la fracture numérique.

Il est fondamental pour l'avenir de nos sociétés du savoir et pour la fracture du développement entre le Nord et le Sud de favoriser un domaine public d'information, en particulier dans les sciences. Tout le monde sait que le cyberespace doit comprendre et refléter la diversité linguistique et culturelle du monde. Mais l'important est de savoir quand cela se produira.

Topologie du contenu

Quelles sont les caractéristiques de ce que l'on appelle communément le « contenu » en ligne? Voyez ces chiffres : le nombre d'hébergeurs internet a dépassé les 500 millions en 2008[1], alors que le nombre des internautes est évalué à environ 1,4 milliards[2], le nombre de sites web à au moins 100 millions et le nombre de pages web visibles[3] à au moins 140 milliards[4].

Tableau 1 : **Statistiques sur l'internet à l'échelle mondial**

Internautes	1,4 milliard
Domaines enregistrés	140 million
Sites web	100 millions - 170 millions*
Pages web	140 milliards - un billion
Pages web indexées	20 milliards - 40 milliard

* Les différences entre les chiffres sont probablement attribuables aux sites virtuels hébergés sur des serveurs. Voir news.netcraft.com/archives/web_server_survey.html

On obtient une idée de la topologie de l' « univers du contenu » en établissant les ratios suivants :

- Trois utilisateurs par hébergeur internet
- Un nom de domaine pour dix utilisateurs
- Un site web pour quatorze utilisateurs
- 1 000 pages web par utilisateur, dont 150 sont indexées par les moteurs de recherche.

Ces ratios sont probablement restés relativement stables au fil des ans, sauf le dernier. Ces dernières années, le pourcentage des pages indexées a diminué à moins de 15 % du total, une tendance qui pourrait rendre les utilisateurs beaucoup plus vulnérables aux diverses distorsions qui conditionnent leur accès au contenu, en plus d'être plus malléables face à des stratégies publicitaires ciblées[5] lancées par les services de recherche.

Distorsions dans l'accès au contenu

Depuis des années, des applications puissantes comme Google suivent notre comportement de navigation sur le web et font ainsi peser une menace sur notre vie privée en ligne. Des données empiriques laissent à penser que l'ordre de présentation des résultats de la recherche est non seulement décidé

1 www.isc.org/index.pl

2 www.internetworldstats.com

3 Le web invisible (également appelé « web profond ») est la somme des pages dynamiques produites par les bases de données ou autres mécanismes programmés qui produisent des pages dynamiques. Certains auteurs estiment qu'il pourrait être 100 à 500 fois plus important que le web visible. Voir Bergman (2001).

4 Il est actuellement impossible de trouver des données sur le nombre total des pages web visibles. Ce chiffre a été extrapolé par l'auteur d'après les chiffres des années précédentes.

5 Jusqu'à présent, la publicité est le principal moteur de l'économie du contenu.

par l'algorithme de classement qui a fait le succès de Google, mais qu'il puise dans l'historique de nos recherches pour nous cibler avec des liens sponsorisés. En outre, des mots clés sont vendus à des intérêts commerciaux, ce qui remet en cause toute l'idée de « l'extraction objective de l'information ». Si on ajoute le fait que 85 % du web visible échappe aux moteurs de recherche[6], la situation commence à ressembler à une forme subtile de censure[7]. Des plaintes commencent à se faire entendre[8] et les gens devraient suivre de près l'évolution qui fait d'une compagnie comme Google un allié de l'accès ouvert, comme avec son initiative Google Scholar[9], mais aussi un opérateur commercialement partial qui se sert de son interface de recherche de base pour réaliser des profits.

Diversité du contenu

Les chiffres moyens cités plus haut sont intéressants pour comprendre l'univers du contenu, mais comme toujours avec les moyennes, ils cachent le facteur de la diversité.

La répartition des internautes entre les régions[10] montre l'Afrique à 4 % et un taux de pénétration de l'internet de seulement 5 %, alors que l'Europe représente 27 % des internautes et un taux de pénétration de 48 %. La répartition des internautes par langue montre l'anglais à 30 %, suivi du chinois (17 %), de l'espagnol (9 %), du japonais (7 %), du français (5 %) et de l'allemand (5 %). Quant à la répartition du contenu sur le web par langue, il n'existe pas de source unique et les chiffres pour l'anglais divergent[11].

Effectivement, la fracture numérique n'est pas seulement une question d'accès : c'est également, et peut-être plus, une question de contenu. Les études de FUNREDES[12] ont montré, par exemple, qu'on produit plus de pages web en français en Grande-Bretagne (0,4 % du total) et en Allemagne (0,5 %) que dans toute l'Afrique (0,3 %), et que la France produit plus de pages en anglais (0,7 % du total) que toute l'Afrique (0,3 %, dont 80% par l'Afrique du Sud). De plus, les tendances observées ne montrent aucune amélioration depuis cinq ans. Les études du Language Observatory Project (LOP)[13] en Asie et en Afrique indiquent que les langues locales représentent un pourcentage de pages web de l'ordre de 1 %, 0,1 % ou 0,01 % par rapport aux langues transfrontalières (anglais, français, russe ou arabe).

La fracture numérique telle qu'elle apparaît à la lumière du contenu semble beaucoup plus profonde que l'écart en matière d'accès : l'Afrique compte 4 % des internautes, alors que les langues africaines représentent moins de 0,4 % du contenu, et les langues africaines moins courantes moins de 0,04 %!

Et ces chiffres s'appuient sur les 5 % de langues dans le monde qui ont une existence numérique – ce qui veut dire qu'il existe un mécanisme de codification pour transcrire leur alphabet sous forme numérique. L'homme a créé quelque 40 000 langues tout au long de son histoire dont 7 000 environ sont toujours utilisées, mais 350 seulement ont une existence numérique.

Pour que l'internet soit une ressource universelle, il ne suffira pas de connecter le monde entier. Il faudra que chacun puisse communiquer sur le net dans sa propre langue maternelle ce qui implique, bien entendu, l'existence d'un contenu équilibré dans la langue de chacun.

Les pragmatiques qui croient que cet objectif est irréaliste – et qu'il est donc acceptable d'obliger les gens à travailler en ligne dans une langue autre que la leur et/ou que l'anglais soit la lingua franca naturelle du cyberespace – devraient réfléchir à ceci : les études de l'UNESCO[14] ont montré que les enfants qui ne sont pas élevés dans une langue maternelle subissent un sérieux handicap. Et les statistiques linguistiques de Wikipedia[15] montrent la présence d'articles dans 264 langues, ce qui prouve bien qu'un autre monde de contenu est possible…

Au Forum sur la gouvernance d'internet (FGI)[16] de novembre 2007, on a commencé à prendre acte des questions de nature linguistique et culturelle, comme en témoigne la table ronde présidée par Gilberto Gil, le ministre de la culture du Brésil de l'époque, et à laquelle le président du Réseau mondial pour la diversité linguistique (RMDL)[17], Adama Samassekou, a fait une intervention. Pourtant ce n'est que la pointe de l'iceberg : même si le système internationalisé des noms de domaines (IDN)[18] fera avancer les choses lorsqu'il

6 Au moment de la rédaction de cet article, un nouveau service est apparu, cuil.com, qui prétend ne pas conserver l'historique de recherche, ce qui est une bonne chose, et indexer pratiquement l'intégralité du web (122 milliards de pages). Malheureusement, les résultats jusqu'ici contredisent ces affirmations.

7 Les études de FUNREDES montrent en particulier que le contenu en anglais est surreprésenté dans les index des moteurs de recherche. Voir l'Observatoire de la diversité linguistique et culturelle dans l'internet : funredes.org/lc.

8 Voir en particulier www.iicm.tugraz.at/iicm_papers/dangers_google.pdf

9 scholar.google.com

10 www.internetworldstats.com/stats.htm

11 Certaines sources prétendent que le pourcentage des pages web en anglais est supérieur à 70% depuis dix ans, malgré l'évolution marquée de la démographie des utilisateurs, alors que d'autres, comme FUNREDES, citent des chiffres de moins de 50 % (funredes.org/lc). Voir UNESCO (2005).

12 Voir les études mentionnées ci-dessus à funredes.org/lc

13 www.language-observatory.org

14 portal.unesco.org/education/en/ev.php-URL_ID=21260&URL_DO=DO_TOPIC&URL_SECTION=201.html

15 en.wikipedia.org/wiki/Wikipedia :Multilingual_statistics

16 Voir www.intgovforum.org et en particulier www.intgovforum.org/Rio_Meeting/IGF2-Diversity-13NOV07.txt

17 Le Réseau mondial pour la diversité linguistique, comme le FGI, est également le produit du Sommet mondial sur la société de l'information (SMSI). Voir www.maaya.org

18 en.wikipedia.org/wiki/Internationalized_domain_name

permettra aux utilisateurs de naviguer sur le web avec des liens écrits dans d'autres jeux de caractères, on sera encore loin de relever le défi du cyberespace qui consiste à refléter la véritable diversité linguistique et culturelle de notre planète.

Au moment de la rédaction de ce rapport, le nombre des internautes en Chine venait de dépasser celui des États-Unis (258 contre 220 millions)[19] – encore un signe de l'accélération du rythme du changement dans la démographie de l'internet. Il s'agit d'une simple question d'inflexion des courbes qui se rapprochent de la saturation lorsque le taux de pénétration dans un segment devient trop élevé (le taux de pénétration de l'internet aux États-Unis est supérieur à 70 % et le chiffre des anglophones qui se connectent est supérieur à 50 %).

Les études de FUNREDES mentionnées plus haut ont montré qu'il y avait initialement un lien entre l'augmentation du nombre des internautes et celle du contenu dans une langue donnée. Mais au fil du temps, on produit moins de contenu par rapport au nombre d'internautes : les nouveaux utilisateurs se comportent davantage comme des consommateurs que des producteurs. Le chaînon manquant est sans doute la compétence numérique, qui comprend également la sensibilisation des utilisateurs à l'importance de la production de contenu.

Contenu ouvert, la grande question au niveau mondial

L'objectif ultime de l'accès universel aux services de télécommunication est de permettre la communication et l'accès à l'information et au savoir. L'autonomisation de tous par le savoir est effectivement au cœur de la société de l'information, et le plus grand domaine public d'information doit alors être considéré comme un droit humain fondamental lié à la cohésion sociale et au développement économique.

L'initiative de Creative Commons[20] offre un large éventail de possibilités pour la protection légale du contenu pour qu'il devienne un contenu ouvert de domaine public, mais cela pose un problème important pour la protection traditionnelle du droit d'auteur. Il s'agit donc d'essayer d'inverser la tendance qu'ont les gens à surprotéger leur contenu et d'encourager une approche plus ouverte dans l'intérêt général sans nuire à un intérêt en particulier.

L'information de domaine public, ou mise en commun de l'information, désigne la liberté d'accéder au travail intellectuel, ou au média sur lequel il est mémorisé, dont l'utilisation n'enfreint aucun droit de propriété intellectuelle ou aucun autre droit communal (comme les droits indigènes) ou toute obligation de confidentialité[21]. La société du savoir doit se construire sur le plus grand domaine public possible pour réaliser son ambition.

Accès ouvert au contenu : Un thème emblématique qui semble progresser

L'accès ouvert renvoie aux publications scientifiques placées dans le domaine public plutôt que d'être conservées par les éditeurs ou les publications. La situation actuelle est la suivante : les fonds publics financent les chercheurs, mais le produit de leurs recherches finit entre les mains d'éditeurs qui accaparent légalement le droit de propriété intellectuelle du chercheur et indirectement de l'administration publique et du contribuable. Cette main mise sert à financer un système d'édition et de publications qui prévoit un examen par les pairs. Celui-ci garantit le prestige d'une publication duquel les chercheurs dépendent pour être reconnus et obtenir des crédits. On voit bien le genre de difficulté que représente un changement du régime du droit d'auteur – de nombreux intérêts sont en jeu !

Ce serait un progrès extraordinaire de la collaboration internationale et un coup de fouet donné à la recherche dans le monde en développement, qui peut difficilement se permettre le prix élevé des publications scientifiques, de voir cette richesse de savoir scientifique librement accessible sur un simple click. Malheureusement, la résistance enracinée dans un système créé pour l'ère de l'imprimerie empêche cette évolution.

Ce n'est pas que le monde scientifique n'ait pas essayé de faire avancer ce dossier, comme en témoigne la Déclaration de Berlin sur le Libre Accès à la Connaissance en Sciences exactes, Sciences de la vie, Sciences humaines et sociales[22] en 2003 et les initiatives comme la Public Library of Science (PLoS)[23], qui offrent des solutions concrètes. La Scholarly Publishing and Academic Resources Coalition (SPARC)[24] élabore des stratégies de plaidoyer à l'appui des politiques publiques sur l'accès ouvert et fait part des progrès.

Encore une fois, le sujet de la diversité linguistique et culturelle n'est pas neutre dans la lutte pour l'accès ouvert, car le système dominant a largement contribué à faire de l'anglais la langue de communication scientifique par excellence. ■

19 Voir Barboza (2008). À noter que les derniers chiffres de Internet World Stats sont différents (Mai 2008 : É.-U. = 220, Chine = 210). Outre le nombre des hébergeurs, les chiffres ne sont pas fiables à 100%.

20 creativecommons.org

21 unesdoc.unesco.org/images/0012/001297/129725f.pdf

22 openaccess.inist.fr/openaccess/spip.php?article38&decoupe_recherche=D%C3%A9claration%20Berlin

23 www.plos.org

24 www.arl.org/sparc

Références

Barboza, D., China Surpasses U.S. in Number of Internet Users, *The New York Times*, 26 juillet 2008. Voir à : www.nytimes.com/2008/07/26/buisiness/worldbusiness/26internet.html?_r=l&oref=slogin

Bergman, M., *The Deep Web : Surfacing Hidden Value*. Ann Arbor : Scholarly Publishing Office, University of Michigan Library, 2001. Voir à : quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0007.104

Creative Commons : fr.creativecommons.org

Déclaration de Berlin sur le Libre Accès à la Connaissance en Sciences exactes, Sciences de la vie, Sciences humaines et sociales : openaccess.inist.fr/openaccess/spip.php?article38&decoupe_recherche=D%C3%A9claration%20Berlin

FUNREDES - Observatoire de la diversité linguistique et culturelle dans l'internet : funredes.org/lc

Forum sur la gouvernance de l'internet : www.intgovforum.org

Internet Systems Consortium : www.isc.org/index.pl

Internet World Stats : www.internetworldstats.com

Netcraft, September 2008 Web Server Survey. Voir à : news.netcraft.com/archives/web_server_survey.html

Public Library of Science (PLoS) : www.plos.org

Réseau mondial pour la diversité linguistique : www.maaya.org

Scholarly Publishing and Academic Resources Coalition (SPARC) : www.arl.org/sparc

UNESCO, Mesurer la diversité linguistique sur Internet. Voir à : portal.unesco.org/ci/fr/ev.php-URL_ID=20804&URL_DO=DO_TOPIC&URL_SECTION=201.html