



# LA PLACE DU FRANÇAIS DANS L'INTERNET

*Étude réalisée pour l'INTIF  
(Institut francophone des nouvelles technologies  
de l'information et de la formation)*

*Décembre 2002*

Association Réseaux & Développement  
(FUNREDES)

<http://funredes.org>  
[contact@funredes.org](mailto:contact@funredes.org)

Avec le soutien de :  
Union Latine  
<http://www.unilat.org>  
[dtil@unilat.org](mailto:dtil@unilat.org)

# TABLE DES MATIÈRES

---

Introduction

Méthodologie

Limites de la méthodologie

Antécédents

La place du français sur la Toile

Choix du moteur pour cette étude

Résultats relatifs

Hypothèses concernant la nouvelle évolution

*Ralentissement de la production de pages web dans les différents secteurs linguistiques étudiés*

*Influence des langues non prises en compte dans l'étude*

*Hausse importante du nombre de pages en anglais*

*Situation liée aux moteurs de recherche*

Résultats absolus 2003

*Relation entre le nombre de locuteurs et leur présence sur la Toile*

*Vitalité de la production d'information des internautes par langue*

Répartition des pages en français par pays

Conclusions

Annexe 1 : Production de pages en français par domaine

Annexe 2 : Production de pages en anglais par domaine

Annexe 3 : Calcul de la proportion de pages web dans le domaine national

Annexe 4 : Recoupement des résultats à l'aide de la fonction "Recherche par pays" de Google

# INTRODUCTION

---

Cette étude comporte deux volets, l'un générique sur l'ensemble des langues latines (et de l'allemand) et l'autre spécifique au français. Le premier vise à mesurer la présence absolue du français, des autres langues latines et de l'allemand sur la Toile ; le second porte sur une approximation de la production de pages web en français entre les différents pays de la Francophonie (ainsi que d'autres pays). Dans les deux volets, des indicateurs sur la société de l'information susceptibles de guider les politiques linguistiques seront proposés. La méthodologie, qui n'est pas explicitée dans ce document mais dont les références sont présentées dans le chapitre suivant, repose sur l'utilisation des moteurs de recherche.

Le premier volet révèle un apparent redressement de l'anglais par rapport au français et à l'ensemble des langues latines. Cette situation, qui contredit toutes les tendances observées depuis le début de nos mesures, semble incohérente par rapport aux données convergentes sur l'amplification de la réduction de la population d'internautes anglophones. L'explication avancée est liée à l'utilisation des moteurs de recherche et sur le constat que l'espace d'indexation de ces moteurs se réduit notablement par rapport à l'espace total de la Toile, ce qui favorise les pages en anglais. Le problème n'est cependant pas seulement de nature méthodologique, étant donné l'importance capitale des moteurs : comment peut-on parler de pages existant sur la Toile si les moteurs de recherche ne peuvent pas les atteindre ? Ce nouveau phénomène mérite d'être pris en compte dans les politiques de soutien linguistique dans le monde virtuel, qui doivent être orientées vers le soutien à la notoriété des pages francophones et pas seulement vers leur production.

Le deuxième volet permet d'établir la production de pages web en français pour chacun des pays francophones et apporte une indication sur la productivité de chacun de ces pays. Cet état des lieux devrait permettre de mesurer les effets et d'orienter les politiques conduites par l'Agence de la Francophonie pour stimuler la production de pages web en français, et de contribuer aux réflexions à partir de données objectives. Les résultats montrent une faiblesse de la productivité en France et traduisent, de façon pathétique, les effets de la fracture numérique dans la présence trop marginale des pages francophones dans les pays du Sud (où le Maroc apparaît comme le pays le plus en avance). Dans ce contexte, il y a lieu d'être encore plus inquiet pour les langues partenaires de la Francophonie, pour lesquelles la méthode ne permet pas, pour le moment, d'établir des mesures.



# MÉTHODOLOGIE

---

Le processus permettant d'obtenir des données sur le web francophone s'appuie sur les travaux réalisés par FUNREDES **depuis 1996** sur la place des langues et cultures latines dans l'Internet (<http://funredes.org/LC/>), avec l'appui méthodologique de l'Union latine pour la partie linguistique. Cette étude permet d'actualiser les résultats de l'étude L5 menée par FUNREDES en 2001. Les détails concernant la méthodologie, ainsi que les résultats obtenus lors de précédentes mesures, sont disponibles sur la page <http://funredes.org/LC/francais/L5/>.

La méthode repose sur l'étude de l'index d'un certain nombre de moteurs de recherche tels que Google ou Alltheweb, ainsi que sur la sélection d'un **échantillon de termes** <sup>1</sup> selon des **critères linguistiques** <sup>2</sup> garantissant la qualité des résultats. Le résultat obtenu en appliquant des **méthodes statistiques classiques** <sup>3</sup> sur les données est une indication <sup>4</sup> du poids relatif des langues latines par rapport à l'anglais. Pour en déduire la valeur absolue du poids des langues latines dans l'Internet, il faut établir une hypothèse sur le poids absolu de l'anglais sur la Toile, ce qui se fait par recouvrements. L'utilisation continue de cette méthode depuis 1998 permet de donner une image de l'**évolution** du poids du français dans l'Internet au cours des 5 dernières années.

À chaque itération de cette étude, il est nécessaire d'analyser le **comportement des moteurs de recherche** tant dans leur façon d'indexer les pages présentes sur l'Internet que dans la manière dont ils gèrent le résultat des requêtes. Une fois cette analyse faite, le moteur qui répond le mieux aux critères de l'étude est sélectionné. Dans certains cas, quand les anomalies détectées répondent de manière cohérente à une logique compréhensible –comme c'est souvent le cas dans la gestion des diacritiques–, les résultats sont redressés.

La méthodologie utilisée pour obtenir la répartition de l'Internet francophone entre les différents pays reprend la procédure établie pour le cas de la langue espagnole en 2001, à l'occasion du Congrès international sur la langue espagnole de Valladolid <sup>5</sup>. La méthode de comptage à partir des mots de l'échantillon linguistique est appliquée sélectivement à **l'intérieur des domaines Internet concernés**. Le moteur de recherche mesure ainsi le nombre d'occurrences des mots de l'échantillon par domaine <sup>6</sup> (.com, .net, .fr, .sn...). Pour chaque mot de l'échantillon, on obtient une répartition par domaine <sup>7</sup>. La moyenne de ces résultats donne une image de la répartition de l'Internet francophone par domaine.



---

<sup>1</sup> Voir l'échantillon sur la page

[http://funredes.org/LC/francais/L5/L5appendix\\_3.html#table\\_15](http://funredes.org/LC/francais/L5/L5appendix_3.html#table_15).

<sup>2</sup> Voir la liste des critères de sélection linguistiques qui permettent de s'approcher d'une signification et d'une portée sémantique équivalentes ainsi que d'éviter les distorsions :

[http://funredes.org/LC/francais/L5/L5appendix\\_7.html](http://funredes.org/LC/francais/L5/L5appendix_7.html).

<sup>3</sup> La méthode de Fischer est appliquée à partir de l'hypothèse d'une distribution normale.

<sup>4</sup> La moyenne des valeurs est établie, ainsi que les "intervalles de confiance", à 90 et 99%.

<sup>5</sup> Document original en espagnol : <http://funredes.org/LC/L5/valladolid.html>.

Document en français présentant les mêmes travaux :

<http://funredes.org/LC/L5/CahiersNumFinal.html>.

<sup>6</sup> 85 domaines sont pris en compte, soit plus de 11 000 recherches gérées automatiquement.

<sup>7</sup> Par exemple : une recherche sur Google pour le mot "vérité" donne 700 000 occurrences. Si on restreint cette recherche au domaine .fr, le comptage est réduit à 202 000, c'est-à-dire 29% de 700 000 ; si on la

# LIMITES DE LA MÉTHODOLOGIE

---

Cette étude se concentre sur **un seul espace** de l'Internet : la Toile. La méthode permet de déterminer, avec une bonne précision statistique, la présence des langues dans l'espace des pages web indexées par les moteurs de recherche. L'extrapolation des résultats vers l'espace entier est d'autant plus risquée que l'espace des pages indexées est un sous-ensemble de taille réduite de l'espace total. Dans tous les cas, il existe un espace de la Toile qui n'est pas indexé par les moteurs de recherche, constitué de pages protégées par mot de passe, de bases de données et d'une partie de pages générées dynamiquement par des programmes comme Java. Quoique cet espace soit riche en information pertinente et pourrait avoir une taille d'un ordre de grandeur supérieur à la partie "visible" de la Toile, il échappe à nos recherches. Enfin, et c'est bien regrettable, nos résultats ne font aucune espèce de distinction en ce qui concerne la nature, **la qualité et la pertinence** des pages web.

Dans des études précédentes, la place des langues dans l'espace des "**groupes de discussion**" (Usenet) avait été mesurée grâce à un moteur spécialisé (DéjàNews). Ce moteur a été repris par Google en 2001 mais les tentatives d'utilisation ont été abandonnées pour cause d'incohérence des résultats. Il n'est pas exclu que nous puissions revenir dans le futur sur cet espace de mesure.

Enfin, les deux espaces les plus riches de l'Internet puisqu'ils touchent à sa partie la plus noble et la plus humaine, à savoir celui des **courriers électroniques** et celui des **communautés virtuelles**, ne sont pas mesurés. En ce qui concerne les communautés virtuelles, il faut cependant noter que la présence de plus en plus fréquente sur le web des mémoires des contributions dans les listes de discussion et l'existence de nombreux systèmes de conférences sur le web permettent d'en tenir indirectement compte.

Une première approximation à la mesure de la **présence des cultures** sur l'Internet a également été réalisée dans le passé. Quoique ayant de très grandes limitations méthodologiques, cette étude a été réalisée, à trois reprises, en juin 1996, en septembre 1998 et en septembre 2001, avec une cohérence dans la méthode, ce qui a permis d'établir d'intéressants paramètres sur les évolutions <sup>8</sup>.

Pour le lecteur curieux ou intéressé par un approfondissement, l'ensemble des résultats obtenus depuis 1996 et des détails sur les méthodologies sont documentés, en toute transparence, sur le site <http://funredes.org/lc>. Nous acceptons avec plaisir et intérêt les critiques, commentaires et suggestions, que vous pouvez adresser à : [contact@funredes.org](mailto:contact@funredes.org).



---

restreint au domaine canadien .ca, on trouve 75 000, soit un peu plus de 10% ; si on l'applique sur le domaine du Maroc (.ma), le résultat est de 2 000, soit environ 0,3%.

<sup>8</sup> Globalement, une progression de plus de 10% des indicateurs culturels latins par rapport à l'anglais entre 1996 et 1998 et de plus de 50% entre 1998 et 2001 a été constatée.

# ANTÉCÉDENTS

---

En juin 1998, FUNREDES a reçu un soutien de l'**Agence de la Francophonie** pour la quatrième édition de son travail d'observation de la place du français et des langues latines dans l'Internet, lequel a été publié en plusieurs langues sur les sites respectifs de l'Agence de la Francophonie (en français), de l'Union latine (dans les différentes langues latines, à l'exception du français) et de FUNREDES (en anglais). Par la suite, FUNREDES a maintenu, avec la **collaboration de l'Union latine**, une veille non systématique sur le thème en procédant à des mesures complémentaires (cinquième étude langue en août 2000 et troisième étude culture en septembre 2001), en améliorant la méthodologie et en ajoutant l'allemand à la liste des langues mesurées.

Depuis la cinquième et dernière étude des langues, FUNREDES tente, sans soutien extérieur, de maintenir une observation plus systématique avec des mesures trimestrielles (<http://funredes.org/LC/L5/ultimas.html>).

En octobre 2001, à l'occasion du deuxième Congrès de la langue espagnole réalisé à Valladolid, FUNREDES a complété sa panoplie méthodologique avec un travail sélectif de mesure de la production de pages web en espagnol, par pays, fournissant ainsi des indicateurs nouveaux et intéressants pour des politiques publiques en la matière. Les résultats de ce travail limité à l'espagnol sont publiés en espagnol dans les Actes du Congrès (<http://funredes.org/lc/L5/valladolid.html>) et en français dans un numéro spécial des Cahiers du numérique de Hermès (<http://funredes.org/lc/L5/CahiersNumFinal.html>).



# LA PLACE DU FRANÇAIS SUR LA TOILE

## Choix du moteur pour cette étude

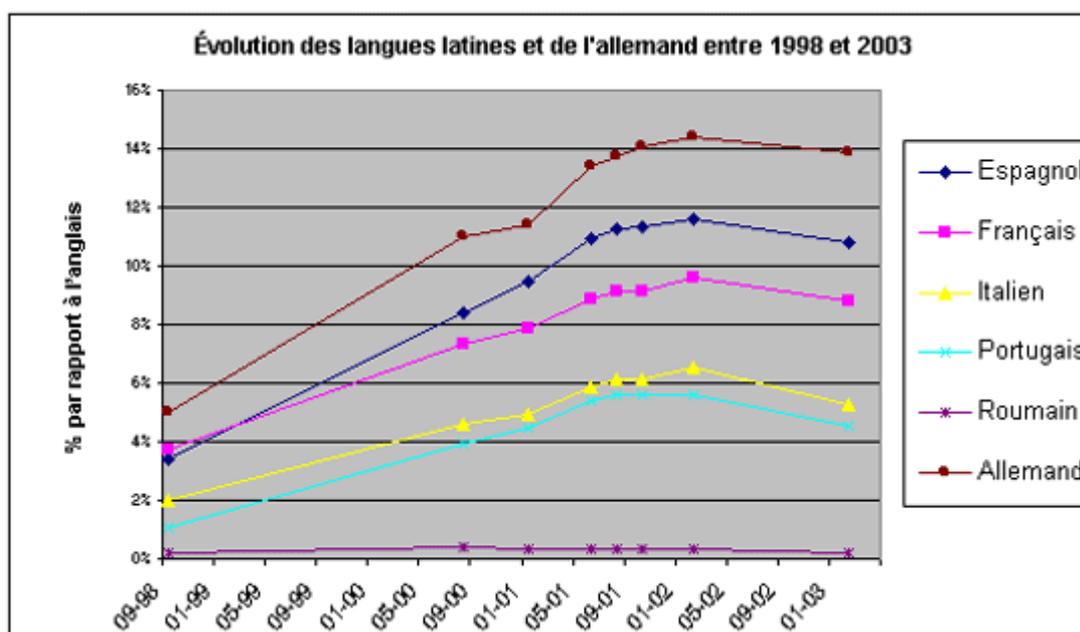
Le moteur de recherche utilisé pour mesurer la place du français est Google. Les raisons de ce choix sont la taille de son index, la cohérence de ses résultats, ainsi que l'amplitude des informations disponibles <sup>9</sup> le concernant. Un atout supplémentaire est celui de sa rapidité de réponse aux requêtes.



## Résultats relatifs

PRÉSENCE RELATIVE SUR LA TOILE PAR RAPPORT À L'ANGLAIS		
	février 2002	décembre 2002
<b>ESPAGNOL</b>	11,60%	10,83%
<b>FRANÇAIS</b>	<b>9,60%</b>	<b>8,82%</b>
<b>ITALIEN</b>	6,51%	5,28%
<b>PORTUGAIS</b>	5,62%	4,55%
<b>ROUMAIN</b>	0,33%	0,23%
<b>ALLEMAND</b>	14,41%	13,87%

On note un net recul des langues latines par rapport à l'anglais par rapport aux valeurs de l'année 2002, et ce pour la première fois depuis le début de nos études. Le graphique suivant présente une vue plus générale de l'évolution du pourcentage relatif des langues latines par rapport à l'anglais au cours des 5 dernières années.



<sup>9</sup> Google Watch par exemple.

L'évolution des langues latines et de l'allemand par rapport à l'anglais est passée par trois phases : une forte augmentation entre 1998 et 2000 <sup>10</sup> suivie d'une stagnation en 2001 et enfin d'une baisse à partir de 2002.



## Hypothèses concernant la nouvelle évolution

Une fois les erreurs possibles dans l'utilisation des moteurs <sup>11</sup> éliminées, le ralentissement et la baisse du poids des langues latines et de l'allemand sur la Toile à partir de 2002 pourraient s'expliquer par plusieurs facteurs (et leur combinaison) :

- Le ralentissement sensible de la production de pages web dans les différents secteurs linguistiques latins et germanophones.
- L'influence indirecte des langues non prises en compte dans l'étude.
- La hausse importante du nombre de pages en anglais.
- Une situation particulière liée aux moteurs de recherche.



### Ralentissement de la production de pages web dans les différents secteurs linguistiques étudiés

Le ralentissement de la production dans les pays de langue latine ou germanique est peu probable. Les chiffres de Global Reach <sup>12</sup> montrent que la proportion d'internautes de langue latine ou germanophone n'a cessé d'augmenter depuis 2001. Le nombre total d'utilisateurs connectés à Internet en 2003 est estimé à 622 millions <sup>13</sup> contre 452 millions en 2001. Dans la même période, la **proportion d'internautes anglophones passe de 47% à 37%**, tandis que la proportion de locuteurs de langues latines ou germaniques passe de 20% à 26%, ce qui implique une réduction de plus de 15% de l'écart. Les résultats précédents de l'étude LC de FUNREDES ont montré une **corrélation quasi linéaire** entre la proportion de pages produites dans une langue donnée et la proportion d'internautes locuteurs de cette langue. La baisse **uniforme** des langues étudiées par rapport à l'anglais renforce la conviction qu'il faut chercher ailleurs la cause de ce nouveau phénomène.



### Influence des langues non prises en compte dans l'étude

Depuis trois ans, le poids des langues non européennes sur l'Internet croît sensiblement. Selon Global Reach, en un an, la proportion d'internautes locuteurs des langues non couvertes par l'étude de FUNREDES est passée de 32% à 37% avec un record pour la population d'internautes chinois qui avoisine maintenant les 70 millions (plus de 10% du total).

Dans la mesure où les **résultats bruts** de l'étude de FUNREDES dérivent d'une **relation entre les langues latines et l'anglais**, la production de pages dans d'autres langues ne

---

<sup>10</sup> On notera que la présence du français passe sous celle de l'espagnol à partir de 1999.

<sup>11</sup> En novembre 2002, ni Google, ni AlltheWeb ne donnaient de résultats consistants (mauvaise gestion des signes diacritiques, résultats aléatoires). En décembre 2002, de nouvelles mesures ont montré que Google donnait à nouveau des résultats cohérents pour notre étude.

<sup>12</sup> <http://www.greach.com/globstats/>. Global Reach mène depuis des années un travail de compilation concernant le nombre d'utilisateurs d'Internet par langue et par pays. Même s'il n'y a pas de consistance dans les sources utilisées pour construire ces indicateurs (auxquels nous donnons une confiance de  $\pm$  20%), cela reste un outil raisonnablement fiable pour comprendre les tendances.

<sup>13</sup> Les résultats du français ont été mis à jour à partir des informations du site <http://www.mediametrie.fr> servant de source à Global Reach.

devrait pas avoir d'influence directe. Cependant, il est plausible que cette population de nouveaux internautes (surtout venue d'Asie) puisse produire en anglais de manière significative <sup>14</sup>, ce qui renforcerait la croissance des pages en anglais et expliquerait donc la baisse (relative) simultanée des résultats des langues latines et de l'allemand.

Cette hypothèse a été **infirmée** par une étude complémentaire de la répartition linguistique de l'Internet en anglais entre les différents domaines concernés (en particulier, .cn et .kr <sup>15</sup>). Il en est ressorti que la production des pages en anglais en dehors des domaines génériques (.com, .net, .edu) et des domaines nationaux des principaux pays anglophones (.us, .uk, .au, .ca) est très faible <sup>16</sup>. Le détail des résultats de ce complément d'étude est disponible en **annexe 2**.



### Hausse importante du nombre de pages en anglais

Un gain relatif de croissance dans la production des pages en anglais par les grands pays anglophones (Etats-Unis, Canada, Angleterre et Australie) serait une explication tout-à-fait plausible dans la mesure où elle se traduirait par une baisse uniforme de la proportion de pages dans les autres langues. Cependant, rien ne permettrait d'expliquer un tel phénomène, surtout dans la mesure où la proportion d'internautes anglophones ne cesse de diminuer. Quoique rien ne permette d'écarter formellement la possibilité d'une augmentation de la production de pages web en anglais, notre analyse nous conduit à favoriser plutôt la dernière hypothèse, celle d'une situation conséquence de l'utilisation des moteurs de recherche dans notre méthodologie.



### Situation liée aux moteurs de recherche

Il n'y a pas de chiffres sûrs concernant le nombre total de pages de la Toile. Parmi les études les plus récentes, Cyveillance estimait, en juillet 2000, que l'Internet contenait plus de 2 milliards de pages et que la croissance était exponentielle. On pourrait estimer, à partir de cette étude et par d'autres recoupements que la taille de la Toile, en 2003, est de l'ordre de **20 milliards de pages**. D'autres études <sup>17</sup> analysent le phénomène du "web invisible" <sup>18</sup> et estimaient, en 2000, le nombre de pages web à plus de 500 milliards.

Le nombre de pages de la Toile est un facteur clé dans l'évaluation du moteur de recherche qui va être utilisé pour l'étude. Si l'on considère les chiffres de Cyveillance et ceux présentés par Google concernant la taille de son index en 2000, on peut dire que Google indexait **la moitié** des pages présentes sur l'Internet à cette date (l'ordre de grandeur était le même pour AltaVista entre 1996 et 1999). On pouvait donc en déduire

---

<sup>14</sup> Par exemple, une présence de l'ordre de 5% des pages en anglais dans le domaine de la chine (.cn) aurait une influence significative compte tenu de la vitesse d'accroissement du domaine chinois. Un tel chiffre serait plausible dans la mesure où les internautes sinophones représentent maintenant plus de 10% de la population mondiale d'internautes.

<sup>15</sup> Chine et Corée.

<sup>16</sup> Par exemple, seulement 0,25% des pages web en anglais appartiennent au domaine .cn.

<sup>17</sup> Celle de BrightPlanet par exemple : <http://www.brightplanet.com/>.

<sup>18</sup> Le "web invisible" est la partie de l'Internet non indexée par les moteurs de recherche et qui, selon BrightPlanet, est plus de 400 fois plus étendue que la partie indexée par les moteurs. Le web invisible comprendrait les pages web qui ne sont pas référencées (aucun lien ne pointant vers ces pages), les pages protégées par un mot de passe, les documents aux formats non indexables, de nombreuses bases de données, ainsi que les réseaux intranet. Pour plus d'informations (en anglais) : <http://www.brightplanet.com/deepcontent/tutorials/DeepWeb/index.asp>.

avec une certaine assurance que l'index des moteurs de recherche donnait une bonne image statistique de la répartition des langues sur la Toile.

Avec une hypothèse de 20 milliards de pages sur l'Internet en 2003, Google n'indexe plus que **15% de l'espace** web visible total. Dans ce contexte, il est clair que les **propriétés statistiques** de l'échantillon indexé ont une forte influence sur nos résultats, puisque ce que nous mesurons objectivement est le pourcentage de pages **indexées** dans une langue donnée par rapport aux pages indexées en anglais <sup>19</sup>.

Le mode d'indexation de Google repose sur le critère de **popularité** d'une page <sup>20</sup>. Le but de ce critère est de favoriser les pages les plus visitées et les plus référencées dans l'ordre d'apparition des réponses aux requêtes. Cependant, l'une de ces conséquences est qu'il **élimine** de l'espace de recherche les pages vers lesquelles le nombre de liens est très faible ou provient de sites eux-mêmes considérés comme peu populaires. Il est clair que cette méthode a tendance à amplifier les écarts dans les deux sens (elle accélère la visibilité des pages qui sont bien référencées et, en même temps, limite l'essor des pages non indexées). Il est également logique que la méthode favorise les pages les plus anciennes –qui ont eu le temps de se faire une niche de popularité– et **pénalise les pages nouvelles**, surtout dans des langues peu répandues car la probabilité de liens sera d'autant plus faible.

Dans la mesure où la communauté d'internautes anglophones est la plus nombreuse et la plus ancienne sur l'Internet, on peut raisonnablement en déduire que les pages en anglais ont une probabilité plus forte d'être présentes lors d'une indexation partielle des pages. Comme les moteurs de recherche indexent, depuis 2001, un pourcentage de plus en plus faible de pages, on peut donc légitimement penser que l'Internet en anglais est de plus en plus favorisé. C'est la situation la plus plausible pour expliquer la baisse relative du poids des langues latines et de l'allemand par rapport à l'anglais dans l'étude de FUNREDES entre 2002 et 2003.

Cette dérive des moteurs de recherche pose bien sûr la **question de la validité** des derniers résultats de cette étude et de la méthodologie employée. Deux facteurs sont à prendre en compte pour répondre à cette question :

- 1) Ce travail reste, encore aujourd'hui (!), **le seul** à produire des chiffres de manière régulière et avec une totale **transparence** sur les méthodes et procédures utilisées.
- 2) Quelle est l'**existence réelle** d'une page non indexée par un moteur ? La vision de la Toile par les moteurs de recherche conditionne implicitement la vision des utilisateurs. Une page non indexée est *virtuellement* inexistante, bien qu'elle existe dans l'*espace virtuel* ☺!

---

<sup>19</sup> Aux personnes qui peuvent s'étonner du déploiement d'un tel arsenal méthodologique alors que les moteurs sont eux-mêmes capables de reconnaître les langues et qu'il est possible, avec un peu d'astuce, de produire leur propre comptage des pages par langues (voir la méthode du complément de l'ensemble vide à l'URL <http://www.funredes.org/LC/francais/L3.html>), nous devons répondre que, de manière consistante, nos études ont montré que ces algorithmes ne sont pas fiables et, en tout cas, pas suffisamment pour une mesure de la place des langues.

<sup>20</sup> La popularité d'une page dépend du nombre d'autres pages qui y font référence (nombre de liens), ainsi que de la popularité du site à partir duquel ces références sont établies, la récurrence indirecte ainsi introduite faisant l'originalité de la méthode.

Les résultats de l'étude FUNREDES/Union latine ne présentent plus la répartition linguistique de la Toile, sinon **la répartition linguistique de la Toile** rendue **visible** par les moteurs de recherche. Ce constat a évidemment de fortes implications sur les politiques efficaces pour les contenus dans une langue donnée, comme cela sera expliqué dans le chapitre de conclusion.



## Résultats absolus 2003

La répartition absolue de l'anglais, des langues latines et de l'allemand sur l'Internet est déterminée à partir de chiffres relatifs obtenus par FUNREDES, d'autres facteurs tels que le nombre d'internautes par langue, un recoupement avec les résultats précédents, ainsi que d'études parallèles. On peut estimer les résultats fiables dans une fourchette de plus ou moins 10%.

Présence absolue sur la Toile	
ANGLAIS	45,0%
ESPAGNOL	4,87%
FRANÇAIS	3,97%
ITALIEN	2,38%
PORTUGAIS	2,05%
ROUMAIN	0,10%
ALLEMAND	6,24%
AUTRES LANGUES	35,39%



### Relation entre le nombre de locuteurs et leur présence sur la Toile

Il est évident que les valeurs de présence absolue ne sont pas un indicateur parfait de la vigueur d'une langue sur les réseaux. Pour obtenir un résultat significatif, il convient de proportionner les valeurs exprimant la présence des langues sur l'Internet à l'aune de leur présence dans le monde réel. La présence **relative** de ces langues est calculée sans tenir pleinement compte du facteur "plurilinguisme". Cette méthode comporte des écueils méthodologiques qui ont été décrits lors de l'étude L4.

*Poids des langues étudiées  
(Source Union latine <sup>21</sup> - chiffres arrondis en millions, 2000)*

	Anglais	Espagnol	Français	Italien	Portugais	Roumain	Allemand
Présence absolue (nombre de locuteurs)	630	375	130	60	190	30	120
Présence relative (pourcentage mondial)	10,50%	6,25%	2,17%	1%	3,17%	0,50%	2%

<sup>21</sup> Voir explications :

[http://www.unilat.org/dtil/lenguainternet/es/lengua/lenguas\\_anexo.htm#Anexo 2](http://www.unilat.org/dtil/lenguainternet/es/lengua/lenguas_anexo.htm#Anexo 2).

### Présence pondérée sur la Toile

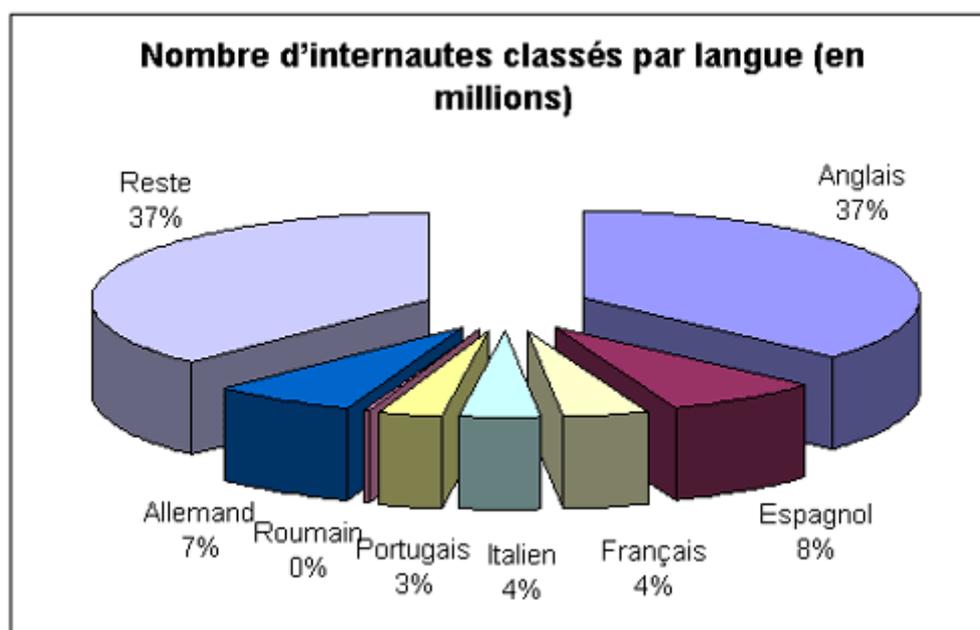
	Présence absolue 2003	Présence pondérée 1998	Présence pondérée 2000	Présence pondérée 2003
ANGLAIS	45%	7,14	5,71	4,29
ESPAGNOL	4,87%	0,40	0,78	0,78
FRANÇAIS	3,97%	1,30	2,02	1,83
ITALIEN	2,38%	1,50	2,77	2,38
PORTUGAIS	2,05%	0,26	0,68	0,65
ROUMAIN	0,10%	0,30	0,38	0,21
ALLEMAND	6,24%	Non disponible	3,15	3,12

Un quotient égal à 1 est à considérer comme un résultat "normal" ; s'il est inférieur à 1, comme *faible* et s'il est supérieur à 1, comme *respectable*.



### Vitalité de la production d'information des internautes par langue

Une étude de Global Reach <sup>22</sup> propose une valeur pour le nombre d'utilisateurs de l'Internet par langue :



En mettant ces résultats en rapport avec ceux de notre étude, il est possible d'en déduire quels sont les segments linguistiques qui produisent le plus d'informations sur la Toile.

<sup>22</sup> <http://www.greach.com/globstats>

### Productivité des locuteurs

	Pages	Internautes	P/I
ANGLAIS	45%	37%	<b>1,22</b>
ESPAGNOL	4,87%	7,6%	0,64
<b>FRANÇAIS</b>	<b>3,97%</b>	<b>4,2%</b>	<b>0,95</b>
ITALIEN	2,38%	3,9%	0,62
PORTUGAIS	2,05%	3,1%	0,67
ROUMAIN	0,10%	0,4%	0,28
ALLEMAND	6,24%	6,8%	<b>0,92</b>

Après l'anglais, le français est la langue la mieux positionnée sur l'Internet par rapport au nombre d'internautes francophones.



## RÉPARTITION DES PAGES EN FRANÇAIS PAR PAYS

Les mesures de la répartition du web en français donnent, pour chaque domaine Internet concerné, les résultats suivants <sup>23</sup>:

Pays <sup>24</sup>	Domaines nationaux	Proportion de pages en français	Domaines génériques	
Allemagne	.de	0,703%	.com	26,120%
Belgique	.be	3,385%	.org	14,333%
Brésil	.br	0,094%	.net	4,573%
Canada	.ca	10,141%	.int	1,047%
<i>dont Québec</i>	<i>.qc.ca</i>	<i>2,048%</i>	.info	0,562%
<i>dont N. Brunswick</i>	<i>.nb.ca</i>	<i>0,015%</i>		
Chine	.cn	0,057%	.edu	0,468%
Côte d'Ivoire	.ci	0,146%	.gov	0,049%
Espagne	.es	0,125%	.tv	0,044%
États-unis	.us	0,042%	.biz	0,020%
France	.fr	30,824%	.coop	0,001%
Italie	.it	0,503%	.aero	0,000%
Japon	.jp	0,053%	.museum	0,000%
Liban	.lb	0,076%	.name	0,000%
Luxembourg	.lu	0,314%	.pro	0,000%
Maroc	.ma	0,168%		
Nlle-Calédonie	.nc	0,069%		
Pays-Bas	.nl	0,126%		
Pologne	.pl	0,057%		
Portugal	.pt	0,057%		
Rep. Tchèque	.cz	0,061%		
Roumanie	.ro	0,070%		
Royaume-uni	.uk	0,285%		
Russie	.ru	0,048%		
Sénégal	.sn	0,049%		
Suisse	.ch	4,241%		
Tunisie	.tn	0,042%		
<i>Autres pays</i>		<i>1,047%</i>		
<b>TOTAL</b>		<b>52,783%</b>		<b>47,217%</b>

<sup>23</sup> Les résultats se lisent de la manière suivante : 30,82% des pages en français appartiennent au domaine .fr.

<sup>24</sup> Seuls les domaines nationaux comprenant plus de 0,04% de l'Internet francophone ont été pris en compte. Le détail pour tous les pays de la Francophonie est en annexe.

Pour établir la **proportion totale** de pages en français pour chaque pays, il convient de répartir le pourcentage de pages produites dans les domaines génériques (c'est-à-dire les 47% de pages en français) entre les différents pays. Pour pouvoir le faire, il faudrait connaître, pour chaque pays, la proportion de pages web hébergées en dehors du domaine national, que l'on peut, pour simplifier, confondre avec **le pourcentage de sites web hébergés dans des serveurs hors domaine national**. Ce pourcentage varie selon le pays considéré <sup>25</sup> et il n'existe pas de données publiques. La plupart des experts consultés –les gestionnaires des domaines nationaux– s'avouent en général incapables de préciser ce pourcentage.

Une étude similaire de la répartition du web en espagnol a pu être réalisée en 2001, notamment grâce à l'aide d'experts nationaux qui ont fourni, de manière solidaire, leurs estimations approximatives concernant l'utilisation du domaine national dans les pays hispanophones. Malheureusement, rares ont été les responsables de domaines francophones qui ont répondu à nos interrogations quand à l'utilisation de leur domaine<sup>26</sup>.

Quand elle n'a pas été donnée par un expert ou estimée par nos soins comme dans le cas des États-unis, la proportion de pages web présentes dans le domaine national de chaque pays a été estimée en tirant profit d'une nouvelle fonction de Google qui permet la recherche d'un mot clé **pour un pays donné**. À l'aide de cette fonction et avec l'astuce de la recherche par "le complément de l'ensemble vide" <sup>27</sup>, il a été possible de connaître le nombre total de pages dans un pays et le nombre total de pages dans le domaine de ce pays, et de construire ainsi le pourcentage souhaité. Les résultats de Google ne sont pas très fiables : nous avons dû les "normaliser" pour combler les lacunes de Google ; mais, quoi qu'il en soit, il s'agit d'une approximation bien meilleure que le chapeau d'un prestidigitateur ! Les détails concernant ces calculs sont disponibles en annexe 3. Bien entendu, nous serions ravis d'intégrer les chiffres que des personnes expertes pourraient à l'avenir nous communiquer, de manière à affiner nos résultats...

Il faut comprendre le chiffre obtenu de la manière suivante : 56% des pages produites en Belgique sont hébergées dans un site du type "www.nom\_du\_site.be".

Le résultat obtenu est une image de la répartition de la **production de pages en français par pays** :

---

<sup>25</sup> En général, lié au schéma de tarification en vigueur (en particulier du différentiel par rapport aux tarifs des domaines génériques) et aux contraintes posées pour la protection des noms de marques.

<sup>26</sup> Une requête a été envoyée par courriel à la liste de distribution des responsables administratifs des noms de domaine ainsi qu'aux chapitres ISOC concernés ; il était précisé que des valeurs très approximatives étaient acceptées sur la base de l'intuition de ces personnes. Seuls MM. Nguyen pour le Viêt-nam et Fuselier pour la Nouvelle-Calédonie nous ont répondu.

<sup>27</sup> On recherche un mot clé comme '-gyewhghedjfgvh' et Google donnera le nombre total de pages indexées qui correspondent au pays donné.

PAYS	Proportion de pages en français dans le domaine national	Proportion de sites web dans le domaine national <sup>28</sup>	Pourcentage de production du total des pages en français
Allemagne	0,703%	88%	<b>0,8%</b>
Belgique	3,385%	56%	<b>6,0%</b>
Brésil	0,094%	89%	<b>0,11%</b>
Canada	10,141%	42%	<b>24,2%</b>
<i>dont Québec</i>	<i>2,048%</i>	<i>42%</i>	<i>4,9%</i>
Chine	0,057%	66%	<b>0,09%</b>
Côte d'Ivoire	0,146%	95%	<b>0,15%</b>
Espagne	0,125%	<i>30%</i>	<b>0,42%</b>
États-unis	0,042%	<i>4%</i>	<b>1,1%</b>
France	30,824%	55%	<b>56,2%</b>
Italie	0,503%	70%	<b>0,07%</b>
Japon	0,053%	84%	<b>0,06%</b>
Liban	0,076%	81%	<b>0,09%</b>
Luxembourg	0,314%	33%	<b>1,0%</b>
Maroc	0,168%	71%	<b>0,24%</b>
Nelle-Calédonie	0,069%	<i>60%</i>	<b>0,12%</b>
Pays-Bas	0,126%	73%	<b>0,17%</b>
Pologne	0,057%	94%	<b>0,06%</b>
Portugal	0,057%	75%	<b>0,08%</b>
Rép. Tchèque	0,061%	91%	<b>0,07%</b>
Roumanie	0,070%	92%	<b>0,08%</b>
Royaume-uni	0,285%	65%	<b>0,44%</b>
Russie	0,048%	89%	<b>0,05%</b>
Sénégal	0,049%	91%	<b>0,05%</b>
Suisse	4,241%	66%	<b>6,5%</b>
Tunisie	0,042%	88%	<b>0,05%</b>
Vietnam	0,004%	<i>93%</i>	<b>0,004%</b>
Autres pays	1,047%	80%	<b>1,3%</b>
<b>TOTAL</b>	<b>52,787%</b>		<b>100%</b>

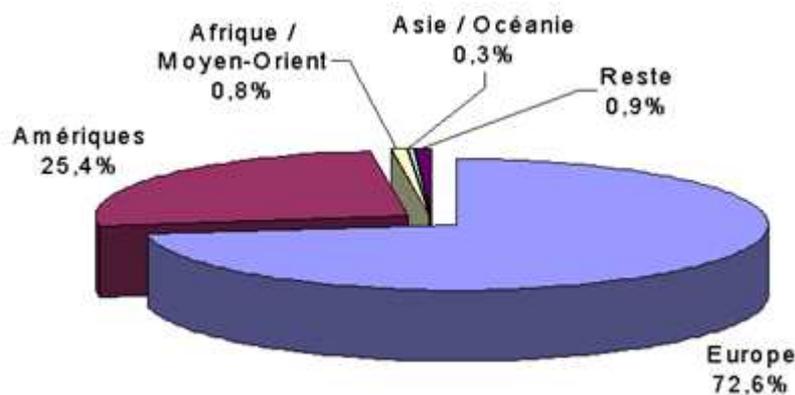
<sup>28</sup> Les chiffres en rouge et italiques sont ceux qui n'ont pas été calculés par la méthode décrite dans l'annexe 3 mais fixés directement à partir d'une estimation d'expert.

On observe que **90% de la production de pages web en français est répartie entre la France, le Canada, la Belgique et la Suisse.**

La production de page en français du Royaume Uni est supérieure à celle de chaque pays d'Afrique francophone. La production totale de l'Afrique en pages francophones est à peine supérieure à celle de l'Italie. Les pays africains les plus producteurs de pages francophones sont, dans l'ordre: le Maroc, la Côte d'Ivoire et le Liban, suivis par le Sénégal et la Tunisie. Les pays suivants produisent une quantité comparable de pages francophones, de l'ordre de 1 pour mille: la République Tchèque, la Pologne, le Portugal, la Roumanie, la Russie, le Brésil, le Liban, le Sénégal, la Chine, le Japon.

Enfin, les États-unis apparaissent comme le premier producteur de pages francophones, après les quatre pays de tête et juste devant le Luxembourg.

La répartition de la production de pages web en français par continent est la suivante :



Plusieurs points ressortent de ce graphique :

- La faible production du continent africain.
- La faible pénétration de la langue française en Asie et en Océanie (région où l'Internet se développe rapidement).
- Le poids énorme de l'Europe et de l'Amérique du nord sur l'Internet en français.

En mettant ces résultats en rapport avec le nombre d'internautes par pays, on obtient les résultats suivants, une sorte de palmarès de la productivité en contenus francophones :

	% Pages en français	Internautes francophones (en millions) <sup>29</sup>	% Internautes francophones	Productivité par pays
<b>Suisse</b>	6,5%	0,90	3,5%	<b>1,88</b>
<b>Belgique</b>	6,0%	1,00	3,8%	<b>1,56</b>
<b>Canada</b>	24,2%	4,40	16,9%	<b>1,43</b>
<b>France</b>	56,2%	18,70	71,9%	<b>0,78</b>
<b>Reste</b>	9,9%	1 <sup>30</sup>	3,85%	
<b>TOTAL</b>	100%	26	100%	1

Ces résultats montrent que la Suisse est la plus productive, suivie de la Belgique et du Canada. Les résultats montrent aussi que les internautes français passent trop de temps à surfer et pas assez à produire des pages en français !



<sup>29</sup> Source : GlobalReach / Médiamétrie.

<sup>30</sup> Le nombre d'internautes francophones dans le reste du monde est incomplet, le nombre présenté (1 million) ne tient compte que des internautes francophones présents aux États-unis et au Luxembourg ; Global Reach ne fournit pas d'informations concernant d'autres pays. Cela, et le fait que des pays non francophones de l'OCDE produisent des proportions non négligeables de pages en français, expliquent la forte productivité du "reste". Dans ces conditions, nous avons préféré ne pas mentionner ce chiffre qui n'est pas fiable.

## CONCLUSION

---

La croissance des langues latines par rapport à l'anglais est, pour la première fois depuis 1996, devenue négative (**perte de 10 à 25%**). Les pourcentages des langues latines par rapport à l'anglais **reculent par rapport aux positions de mi-2001**. L'explication de ce recul n'est à trouver ni dans la réalité d'une baisse de la production des langues latines, ni dans une hausse de la production des pages en anglais. Elle est probablement liée au fait que les moteurs de recherche ne pouvant plus indexer une proportion de pages existantes aussi large que les années précédentes –la proportion des pages indexées par rapport au total passant de 30-50% à 10-15%–, la sélection des pages indexées, basée sur le nombre et la qualité des liens vers une page, favorise l'anglais au détriment des langues latines.

Le **français** est aujourd'hui présent dans environ **4% des pages sur la Toile**. Cela continue de le placer comme une des langues les mieux représentées en proportion du nombre d'internautes de cette langue.

À la lumière des faits précédents, une bonne politique pourrait s'orienter vers la promotion des pages existantes plutôt que seulement vers la production de pages nouvelles, ainsi que vers le soutien aux moteurs francophones et à toutes les actions de promotion de contenus francophones de bonne qualité sur l'Internet. Par ailleurs, il semble que la France ait besoin, spécifiquement au sein de la Francophonie, d'une politique de sensibilisation et d'incitation à la production.

Il est illusoire d'espérer une hausse substantielle de la production des pages francophones en Afrique (et en Asie) et un impact significatif sur la production globale tant que le nombre d'internautes restera d'un ordre de grandeur inférieur... Ici, la politique efficace doit se concentrer sur la **réduction du fossé numérique** en l'accompagnant d'une **éducation** pour obtenir des internautes conscients des enjeux et capables de **produire des contenus**.

Le français, l'espagnol et le portugais ont un espace de locuteurs qui n'est pas limité à un seul pays et joue souvent un rôle de **langue véhiculaire** dans certaines régions du monde. Ce sont des atouts stratégiques pour le monde virtuel globalisé et l'espagnol, en particulier, connaît un essor spectaculaire en raison de sa position géostratégique qui en fait à la fois la langue d'un continent et la seconde langue des Etats-Unis.

Mais qu'en est-il des **langues partenaires** de la Francophonie et des langues indigènes de l'Amérique latine dont certaines comme le swahili ou le quechua jouent également dans leur région un rôle de langue véhiculaire ? Un **effort concerté** est nécessaire pour aider ces langues à trouver leur place dans le cyberspace. Cela passe bien sûr par des actions intelligentes d'**accompagnement à l'accès** : jeux de caractères informatiques pour donner une existence informatique à ces langues, formations à la création de contenus et sensibilisation aux enjeux et, là aussi, moteurs de recherche, comme l'exemple du swahili dans Google <sup>31</sup> rapporté par le magazine Thot. À condition toutefois de bien prendre garde, comme l'indique l'auteur de l'article référencé dans Thot <sup>32</sup>, que *"ces développements soient effectués par les africains eux-mêmes pour que les contenus conviennent à leurs besoins et que ne leur soient pas imposées des idées étrangères"*.



---

<sup>31</sup> <http://www.google.com/intl/sw/>

<sup>32</sup> Voir <http://thot.cursus.edu/rubrique.asp?no=18449>.

# ANNEXES

## Annexe 1 : Production de pages en français par domaine

DOMAINES GÉNÉRIQUES		
	.aero	0,000%
	.biz	0,020%
	.com	26,120%
	.coop	0,001%
	.edu	0,468%
	.gov	0,049%
	.info	0,562%
	.int	1,047%
	.museum	0,000%
	.name	0,000%
	.net	4,573%
	.org	14,333%
	.pro	0,000%
	.tv	0,044%
TOTAL		47,217%

ASIE ET OCÉANIE		
CAMBODGE	.kh	0,001%
CHINE	.cn	0,057%
JAPON	.jp	0,053%
LAOS	.la	0,000%
MAURICE	.mu	0,017%
NLLE-CALÉDONIE	.nc	0,069%
POL. FRANC.	.pf	0,020%
SEYCHELLES	.sc	0,000%
VANUATU	.vu	0,000%
VIETNAM	.vn	0,004%
WALLIS & FUTUNA	.wf	0,000%
TOTAL		0,221%

AMÉRIQUES		
ARGENTINE	.ar	0,022%
BRÉSIL	.br	0,094%
CANADA	.ca	10,141%
CHILI	.cl	0,024%
DOMINIQUE	.dm	0,000%
ÉTATS-UNIS	.us	0,042%
GUADELOUPE	.gp	0,002%
GUYANE	.gf	0,003%
HAITI	.ht	0,000%
MARTINIQUE	.mq	0,003%
NW BRUNSWICK	.br.ca	0,000%
QUÉBEC	.qc.ca	2,048%
ST P. & MIQUELON	.pm	0,000%
STE LUCIE	.lc	0,000%
TOTAL		10,331%

<b>TOTAL GÉNÉRAL</b>	<b>99,247%</b>
----------------------	----------------

<b>Hors Francophonie</b>	<b>36,510%</b>
--------------------------	----------------

EUROPE		
ALBANIE	.al	0,000%
ALLEMAGNE	.de	0,703%
BELGIQUE	.be	3,385%
BULGARIE	.bg	0,007%
ESPAGNE	.es	0,125%
FRANCE	.fr	30,824%
ITALIE	.it	0,503%
LITHUANIE	.lt	0,004%
LUXEMBOURG	.lu	0,314%
MACÉDOINE	.mk	0,002%
MOLDAVIE	.md	0,002%
MONACO	.mc	0,028%
PAYS BAS	.nl	0,126%
POLOGNE	.pl	0,057%
PORTUGAL	.pt	0,057%
REP. TCHEQUE	.cz	0,061%
ROUMANIE	.ro	0,070%
ROYAUME UNI	.uk	0,285%
RUSSIE	.ru	0,048%
SLOVENIE	.si	0,004%
SUISSE	.ch	4,241%
TOTAL		40,846%

AFRIQUE ET MOYEN-ORIENT		
BÉNIN	.bj	0,004%
BURKINA	.bf	0,030%
BURUNDI	.bi	0,014%
CAMEROUN	.cm	0,025%
CAP-VERT	.cv	0,000%
CENTRAFRIQUE	.cf	0,000%
COMORES	.km	0,000%
CÔTE D'IVOIRE	.ci	0,146%
DJIBOUTI	.dj	0,005%
EGYPTE	.eg	0,015%
GABON	.ga	0,003%
GUINÉE	.gn	0,001%
GUINÉE EQUAT.	.gq	0,000%
LIBAN	.lb	0,076%
MADAGASCAR	.mg	0,030%
MALI	.ml	0,006%
MAROC	.ma	0,168%
NIGER	.ne	0,010%
RÉP. DÉM. CONGO	.cd	0,001%
RÉP. CONGO	.cg	0,003%
RÉUNION	.re	0,000%
RWANDA	.rw	0,002%
SÉNÉGAL	.sn	0,049%
TCHAD	.td	0,000%
TOGO	.tg	0,002%
TUNISIE	.tn	0,042%
TOTAL		0,632%



## Annexe 2 : Production de pages en anglais par domaine

DOMAINES GÉNÉRIQUES		
	.com	36,492%
	.org	18,821%
	.edu	13,759%
	.net	5,363%
	.gov	2,572%
	.int	0,224%
TOTAL		77,231%

PAYS ANGLOPHONES		
ROYAUME-UNI	.uk	5,788%
CANADA	.ca	2,575%
AUSTRALIE	.au	2,435%
ÉTATS-UNIS	.us	2,058%
NLLE ZÉLANDE	.nz	0,410%
AFR. DU SUD	.za	0,375%
IRLANDE	.ie	0,339%
HONG KONG	.hk	0,154%
SINGAPOUR	.sg	0,137%
INDE	.in	0,110%
ZIMBABWE	.zw	0,009%
JAMAÏQUE	.jm	0,003%
TOTAL		14,390%

AUTRES DOMAINES		
ALLEMAGNE	.de	1,837%
JAPON	.jp	0,718%
RUSSIE	.ru	0,596%
HOLLANDE	.nl	0,586%
FRANCE	.fr	0,531%
ITALIE	.it	0,438%
SUÈDE	.se	0,435%
CORÉE	.kr	0,270%
CHINE	.cn	0,246%
FINLANDE	.fi	0,228%
NORVÈGE	.no	0,217%
ESPAGNE	.es	0,194%
ISRAËL	.il	0,118%
E.A.U	.ae	0,008%
PORTUGAL	.pr	0,000%
TOTAL		6,422%

<b>TOTAL</b>	<b>98,043%</b>
--------------	----------------



## Annexe 3 : Calcul de la proportion de pages web dans le domaine national

Pour établir la **proportion totale** de pages en français pour chaque pays, il faut additionner les pages comptées dans le domaine national de chaque pays et les pages correspondant à ces pays hébergées dans des serveurs utilisant des noms de domaine génériques <sup>33</sup>. La deuxième partie de la somme pose des problèmes de définition et est particulièrement difficile à établir.

On peut considérer comme une définition acceptable qu'un site web est "**présent**" dans **un pays** s'il est hébergé par un serveur localisé sur son territoire. Avec cette définition, le site d'une entreprise française hébergé aux États-unis sera comptabilisé comme site nord-américain et le site d'une organisation sénégalaise hébergé en France devra être comptabilisé comme français. De même, le site d'un organisme international comme <http://www.unesco.org> ayant son serveur en France sera considéré comme français. On voit bien les limites de la définition.

Pour procéder à la détection de la localité des serveurs, trois types d'information sont disponibles :

- quel est le propriétaire du nom de domaine (fonction whois <sup>34</sup>) ?
- quels sont les serveurs qui donne le service de gestion du nom de domaine (également fourni par la fonction whois) ?
- et quelle est l'adresse IP <sup>35</sup> du serveur ?

Les deux premières informations donnent une idée de l'endroit où peut se trouver le serveur hébergeant un site mais ne permettent toutefois pas d'obtenir une certitude (rien n'interdit d'avoir un propriétaire de site ou un serveur de domaine dans un pays différent de celui du site). Des organismes régionaux (comme arin.net) allouent les numéros d'IP par pays et par fournisseurs. Il est raisonnable de penser qu'ils maintiennent une comptabilité des tranches de numéros d'IP attribués (un peu comme une banque centrale garde les numéros des billets de banque) avec leur destinataire.

Lorsque le moteur Google procède, indépendamment du nom de domaine, à l'identification des pages par pays <sup>36</sup>, il doit forcément faire appel à l'une des trois informations mentionnées. Nous faisons l'hypothèse qu'il le fait au moyen d'une base de données des numéros d'IP qui fait correspondre un pays à un numéro d'IP.

Cette méthode a ses limites. L'observation montre que plusieurs sites ne sont pas clairement associés à un pays (c'est le cas de notre site <funredes.org>). Dans de nombreux cas, les États-unis apparaissent comme le pays de sites extérieurs (c'est le cas du site haïtien <rehred-haiti.net>). L'hypothèse de la base de données (numéro d'IP, pays) permet de comprendre ces anomalies. Un fournisseur d'un petit pays acquiert

---

<sup>33</sup> C'est-à-dire, par exemple, le nombre de pages présentes sur le territoire français dans des sites du type "http://www.mon\_site.com", en plus des pages dans des sites du type "http://www.mon\_site.fr".

<sup>34</sup> Des sites comme celui de "Network Solutions" (<http://www.networksolutions.com/>) donnent des informations sur le propriétaire de différents sites.

<sup>35</sup> L'adresse dans le Protocole Internet : voir <http://www.commentcamarche.net/internet/ip.php3> pour la définition d'une adresse IP.

<sup>36</sup> Cette option est accessible à partir de la section "Outils linguistiques" de Google.

souvent ses séries de numéros d'IP d'un fournisseur des États-unis qui ne lui remet pas forcément des séries identifiables dans la base de données. Par ailleurs, l'amplitude des résultats montre que Google ne détiendrait pour ses recherches qu'environ 10% de cette base de données (en d'autres termes, 90% des sites échappent à la recherche par pays de Google).

Compte tenu de ces limitations, la meilleure option reste de demander l'avis de spécialistes de la gestion des domaines dans chaque pays et de prendre leur estimation, même si elle est très intuitive et donc approximative. En l'absence de réponse, la seule option pour éviter d'indiquer des chiffres arbitraires est de reconstruire les valeurs à l'aide de la fonction de recherche par pays de Google en procédant à un réajustement des chiffres pour les normaliser (faire en sorte que le total cadre et répartir les augmentations de manière cohérente) et en faisant l'hypothèse (laquelle est heureusement confortée par la cohérence des résultats obtenus) qu'il n'y a pas trop de déformation statistique et que l'on peut extrapoler les valeurs obtenues à partir de 10% des sites.

C'est donc ainsi qu'à partir de l'**organisation par pays** de l'index de Google, la proportion de pages web dans un domaine national a été calculée. La technique de complément de l'ensemble vide a été utilisée pour chaque domaine national <sup>37</sup>, puis pour les domaines génériques dans chaque pays <sup>38</sup>. Par exemple, pour la France, on obtient les résultats suivants:

	Domaine national (fr)	.com	.org	.net	.edu	.info	Autres
Nombre de pages (en milliers)	8 740	5 550	3 200	2 880	26	248	149

Il y aurait donc en France 8,7 millions de pages indexées dans le domaine national et 12 millions dans des domaines génériques. On peut donc calculer que, selon Google, 42% <sup>39</sup> des pages présentes en France appartiennent au domaine national.

Les résultats pour l'ensemble des pays sont les suivants <sup>40</sup>:

<sup>37</sup> Par exemple, une recherche du type "-dasfsdafasdfsdfadbb site:.fr" donnera le nombre total de pages présentes dans l'index de Google sous le domaine .fr.

<sup>38</sup> Il faut réaliser, dans la section "Outils linguistiques" de Google, une recherche du type "-dasfsdafasdfsdfadbb site:.DOM" où DOM prend les valeurs des domaines génériques et de pays.

<sup>39</sup>  $8,7/(8,7+12)$

<sup>40</sup> Les quantités de pages sont en milliers.

EUROPE			
	.NAT	.generique	%NAT
ALBANIE	16	0	100%
ALLEMAGNE	40700	10540	79%
BELGIQUE	2710	3524	43%
BULGARIE	384	686	36%
ESPAGNE	3780	7042	35%
FRANCE	8740	12053	42%
ITALIE	8970	6864	57%
LITUANIE	927	74	93%
LUXEMBOURG	232	761	23%
MACÉDOINE	162	10	94%
MOLDAVIE	127	18	88%
MONACO	35	46	43%
PAYS-BAS	10600	6983	60%
POLOGNE	7260	974	88%
PORTUGAL	1300	758	63%
RÉP. TCHÈQUE	6120	1176	84%
ROUMANIE	1400	243	85%
ROYAUME-UNI	18000	16632	52%
RUSSIE	14500	3384	81%
SLOVÉNIE	439	957	31%
SUISSE	6220	5613	53%
TOTAL	132622	78338	63%

AMÉRIQUES			
	.NAT	.generique	%NAT
ARGENTINE	1640	1668	50%
BRÉSIL	7170	1632	81%
CANADA	6520	14730	31%
CHILI	766	342	69%
DOMINIQUE	2	0	100%
ÉTATS-UNIS	6560	49078	12%
GUADELOUPE	4	6,8	36%
GUYANE	3	0	100%
HAÏTI	0	0	
MARTINIQUE	3	1,4	69%
ST P. & MIQUELON	0	0	
STE LUCIE	6	0	100%
TOTAL	22673	67458,2	25%

ASIE ET OCÉANIE			
	.NAT	.generique	%NAT
CAMBODGE	14,70	2	87%
CHINE	6530,00	5876	53%
JAPON	24500,00	8669	74%
LAOS	14,40	0	100%
MAURICE	52,40	9	85%
NLLE-CALÉDONIE	46,90	36	56%
POL. FRANÇAISE	47,40	2	95%
SEYCHELLES	4,66	0	100%
VANUATU	95,60	0	100%
VIETNAM	175,00	47	79%
W. ET FUTUNA	0,01	0	100%
TOTAL	31481,07	14643	68%

AFRIQUE ET MOYEN-ORIENT			
	.NAT	.generique	%NAT
BÉNIN	3,80	3,22	54%
BURKINA	14,60	3,77	79%
BURUNDI	3,41	0,00	100%
CAMEROUN	15,30	1,86	89%
CAP-VERT	3,15	0,00	100%
CENTRAFRIQUE	0,76	0,00	100%
COMORES	0,05	0,00	100%
CÔTE IVOIRE	40,80	4,13	91%
DJIBOUTI	8,71	0,00	100%
EGYPTE	76,80	201,06	28%
GABON	1,08	9,73	10%
GUINÉE	0,95	4,42	18%
GUINÉE EQUAT.	0,00	0,00	
LIBAN	69,70	30,00	70%
MADAGASCAR	14,50	0,07	100%
MALI	4,05	0,74	85%
MAROC	57,70	40,84	59%
NIGER	5,34	0,00	100%
RÉP. CONGO	0,98	0,00	100%
RÉP. DÉM. CONGO	40,00	0,00	100%
RÉUNION	0,45	0,00	100%
RWANDA	4,33	0,95	82%
SÉNÉGAL	37,50	7,25	84%
TCHAD	0,45	0,00	100%
TOGO	2,97	4,41	40%
TUNISIE	55,90	14,54	79%
TOTAL	463,28	327	59%

	.NAT	.generique
<b>TOTAL</b>	<b>187240</b>	<b>160766</b>

Tous les pays de la Francophonie, ainsi que la plupart des pays producteurs de pages web, sont présents dans ce tableau. Comme on le constate, le nombre total de pages est proche de 350 millions alors que l'index de Google comporte plus de 3 milliards de pages indexées. On peut toutefois espérer que cette erreur dans l'indexation des pages

par pays soit uniforme et puisse donner une image crédible de la proportion de pages web par domaine national (%NAT).

Une autre anomalie de ces résultats est clairement celle du chiffre des États-unis que nous n'avons donc pas utilisé. Nous avons multiplié par 2 le chiffre de 2%, que nous avons déterminé lors de l'étude de 2001, pour tenir compte des progrès du domaine .us.

En utilisant la proportion de pages web dans un domaine national avec les résultats calculés pour chaque domaine, on obtient le résultat suivant :

- Avant répartition des domaines génériques :

Domaines génériques	47,2%
Europe	41,2%
Amériques	12,4%
Afrique et Moyen-Orient	0,6%
Asie et Océanie	0,3%
Reste	0,7%
TOTAL	100%

- Après répartition des domaines génériques :

Domaines génériques	0%
Europe	93,9%
Amériques	33,6%
Afrique et Moyen-Orient	0,9%
Asie et Océanie	0,3%
Reste	1,4%
<i>TOTAL</i>	<i>130%</i>

Les pourcentages de sites hors domaine national devront être "normalisés" pour faire revenir le total à 100% <sup>41</sup>. Les résultats obtenus sont ceux présentés dans le tableau en page 12.



---

<sup>41</sup> L'équation  $Tx(x-1)$  est appliquée où  $T = 1,527$ . Il s'agit d'une augmentation non uniforme (en forme de parabole) des valeurs pour combler les 30%, faite de sorte à préserver les valeurs 0% et 100%.

## Annexe 4 : Recoupement des résultats à l'aide de la fonction "recherche par pays" de Google

Malgré les limites de la méthode de recherche par pays de Google, nous n'avons pas pu résister à procéder à une mesure de la répartition des pages francophones par pays en l'utilisant. Pour faire cela, une mesure directe de l'échantillon linguistique a été faite par pays <sup>42</sup> (au lieu de "par domaine"). Cette deuxième méthode est beaucoup moins fiable que la méthode que nous avons développée ; néanmoins, il serait satisfaisant pour l'esprit de constater que deux méthodes totalement différentes conduisent à des écarts pas trop importants (ou bien s'il y a des écarts, d'être capable de comprendre pourquoi). Les résultats sont extrêmement satisfaisants, comme le montre le tableau suivant :

Écart remarquablement faible
Écart normal
Écart fort en %
<b>Écart anormalement fort en %</b>
Écart notable en valeur absolue
Écart anormal en valeur absolue

PAYS	Méthode Google normalisée	Méthode LC	Différence %	Différence simple
ALBANIE	0,000%	0,000%		0,000%
ALLEMAGNE	1,417%	0,802%	-77%	-0,615%
BELGIQUE	4,543%	6,058%	25%	1,515%
BULGARIE	0,006%	0,015%	61%	0,009%
ESPAGNE	1,210%	0,269%	-349%	-0,940%
FRANCE	56,035%	56,727%	1%	0,692%
ITALIE	1,111%	0,728%	-53%	-0,382%
LITUANIE	0,006%	0,004%	-38%	-0,002%
LUXEMBOURG	0,830%	0,969%	14%	0,139%
MACÉDOINE	0,002%	0,002%	-11%	0,000%
MOLDAVIE	0,001%	0,002%	46%	0,001%
MONACO	0,032%	0,050%	36%	0,018%
PAYS-BAS	0,394%	0,174%	-127%	-0,220%
POLOGNE	0,063%	0,061%	-4%	-0,002%
PORTUGAL	0,075%	0,076%	2%	0,001%
RÉP. TCHÈQUE	0,079%	0,067%	-18%	-0,012%
ROUMANIE	0,082%	0,076%	-7%	-0,005%

<sup>42</sup> Au lieu de chercher le nombre de pages pour chaque mot de l'échantillon par domaine (recherche du type "mot\_de\_l'échantillon site:.fr"), une mesure du type "mot\_de\_l'échantillon" avec l'option pays France a été réalisée.

ROYAUME-UNI	1,297%	0,441%	-194%	-0,856%
RUSSIE	0,074%	0,054%	-36%	-0,020%
SLOVÉNIE	0,006%	0,009%	39%	0,004%
SUISSE	6,127%	6,509%	6%	0,382%
<b>EUROPE</b>	<b>73,390%</b>	<b>73,095%</b>	<b>0%</b>	<b>-0,295%</b>

PAYS	Méthode Google normalisée	Méthode LC	Différence %	Différence simple
ARGENTINE	0,030%	0,035%	15%	0,005%
BRÉSIL	0,118%	0,106%	-12%	-0,013%
CANADA	15,059%	24,482%	38%	9,422%
CHILI	0,016%	0,030%	46%	0,014%
DOMINIQUE	0,000%	0,000%		
ÉTATS-UNIS	10,310%	0,246%	-4083%	-10,063%
GUADELOUPE	0,006%	0,004%	-36%	-0,002%
GUYANE	0,002%	0,003%	23%	0,001%
HAÏTI	0,000%	0,000%		0,000%
MARTINIQUE	0,005%	0,004%	-22%	-0,001%
St P. & MIQUELON	0,000%	0,000%		0,000%
STE LUCIE	0,000%	0,000%		0,000%
<b>AMÉRIQUES</b>	<b>25,546%</b>	<b>24,9%</b>	<b>-3%</b>	<b>-0,636%</b>

PAYS	Méthode Google normalisée	Méthode LC	Différence %	Différence simple
BURKINA	0,041%	0,034%	-21%	-0,007%
BURUNDI	0,018%	0,014%	-31%	-0,004%
BÉNIN	0,014%	0,006%	-130%	-0,008%
RÉP. DÉM. CONGO	0,001%	0,001%	-15%	0,000%
CENTRAFRIQUE	0,000%	0,000%		0,000%
RÉP. CONGO	0,003%	0,003%	-15%	0,000%
CÔTE IVOIRE	0,156%	0,154%	-2%	-0,003%
CAMEROUN	0,032%	0,027%	-21%	-0,006%
CAP-VERT	0,000%	0,000%		0,000%
DJIBOUTI	0,007%	0,005%	-38%	-0,002%
EGYPTE	0,023%	0,040%	42%	0,017%
GABON	0,028%	0,021%	-34%	-0,007%
GUINÉE	0,018%	0,004%	-361%	-0,014%
GUINÉE EQU.	0,000%	0,000%		0,000%
COMORES	0,000%	0,000%		0,000%

LIBAN	0,077%	0,094%	18%	0,017%
MAROC	0,244%	0,237%	-3%	-0,007%
MADAGASCAR	0,031%	0,030%	-3%	-0,001%
MALI	0,008%	0,007%	-22%	-0,001%
NIGER	0,013%	0,010%	-27%	-0,003%
RÉUNION	0,000%	0,000%		0,000%
RWANDA	0,002%	0,002%	-3%	0,000%
SÉNÉGAL	0,074%	0,054%	-36%	-0,020%
TCHAD	0,000%	0,000%		0,000%
TOGO	0,010%	0,004%	-171%	-0,007%
TUNISIE	0,061%	0,048%	-27%	-0,013%
<b>AFRIQUE</b>	<b>0,863%</b>	<b>0,794%</b>	<b>-9%</b>	<b>-0,068%</b>

PAYS	Méthode Google normalisée	Méthode LC	Différence %	Différence simple
CHINE	0,086%	0,087%	1%	0,001%
JAPON	0,097%	0,063%	-52%	-0,033%
CAMBODGE	0,007%	0,001%	-541%	-0,006%
LAOS	0,000%	0,000%		0,000%
MAURICE	0,026%	0,019%	-42%	-0,008%
NLLE-CALÉDONIE	0,128%	0,100%	-27%	-0,027%
POL. FRANÇAISE	0,023%	0,021%	-12%	-0,002%
SEYCHELLES	0,000%	0,000%		0,000%
VIETNAM	0,008%	0,005%	-75%	-0,003%
VANUATU	0,000%	0,000%		0,000%
WALLIS & FUTUNA	0,000%	0,000%		0,000%
<b>ASIE/OCÉANIE</b>	<b>0,375%</b>	<b>0,296%</b>	<b>-27%</b>	<b>-0,079%</b>

<b>TOTAL</b>	<b>100%</b>	<b>99%</b>	<b>-1%</b>	<b>-1,078%</b>
--------------	-------------	------------	------------	----------------

Les chiffres en rouge et gras peuvent trouver leur explication dans le fait que ce sont :

- soit des pays exportateurs de numéro d'IP (leurs fournisseurs Internet servent des numéros IP en dehors de leur espace géographique) comme cela doit être le cas, en premier lieu des États-unis et, en moindre proportion, du Royaume-uni et des Pays-Bas.
- soit des pays comme l'Espagne, qui ont une faible utilisation de leur domaine national.

Les chiffres en violet et soulignés ont la même explication sur le côté exportateur de certains pays. Il est remarquable de noter que le trop plein des États-unis correspond au manque du Canada ! La différence d'environ 10% correspondrait-elle aux pages canadiennes francophones logées dans des serveurs des États-unis avec des numéros d'IP non identifiés comme canadiens ?

En tout cas, les chiffres sont remarquablement proches et apportent un nouvel argument de crédibilité à notre méthode.

À noter qu'Isidro F. Aguillo, chercheur du CINDOC en Espagne et responsable de la revue *Cybermetrics*<sup>43</sup>, utilise déjà ces facilités offertes par Google pour des études de création d'indicateurs de la société de l'information comme, par exemple, dans son étude "Indicadores de contenidos para la Web Academica Latinoamericana".



---

<sup>43</sup> <http://www.cindoc.csic.es/cybermetrics/cybermetrics.html>