Contenidos del buscador *Google*. Distribución por países, dominios e idiomas

Por **Isidro F. Aguillo**, **José L. Ortega** y **Begoña Granadino**

Resumen: *Se han analizado de forma cuantitativa los contenidos de la base de datos de Google para recoger el impacto de la nueva infraestructura BigDaddy. Los resultados muestran un incremento sustancial del volumen de páginas indizadas que se acercaría a los 40.000 millones, de los que el 40% pertenecen al dominio .com. El dominio .es representa menos de la mitad de las páginas de España, debido al gran número de páginas registradas bajo dominios internacionales. En total, el peso de nuestro país en los contenidos de Google es de alrededor del 1,1%. Esto es coherente con los resultados obtenidos respecto a idiomas, pues el español ocupa un retrasado sexto puesto (2,6%) y el catalán representa el 0,09% del total. El español es abrumadoramente el idioma principal en los países hispanohablantes, pero su presencia en EUA es en términos porcentuales muy reducida.*

Palabras clave: *Google, Cibermetría, Tamaño, Dominios, Países, Idiomas.*

Title: Contents of the Google database: distribution by country, domain and language

Abstract: *Quantitative analysis of the contents of Google's database in order to ascertain the impact of the new BigDaddy infrastructure. The results show a substantial increase in the volume of pages indexed, approaching 40.000.000.000, of which 40% are in the .com domain. The .es domain represents less than half of the pages from Spain, due to the large number of pages registered under international domains. In all, our country's presence in Google content is approximately 1.1%. This is consistent with results obtained regarding languages, since Spanish occupies a very humble sixth place (2.6%) and Catalan represents 0.09% of the total. Spanish is overwhelmingly the principal language in Spanish–speaking countries, but its presence in the United States is, percentage–wise, quite low.*

Keywords: *Google, Cybermetrics, Size, Domains, Countries, Languages.*

Aguillo, Isidro F.; **Ortega, José L.**; **Granadino, Begoña**. «Contenidos del buscador Google. Distribución por países, dominios e idiomas». En: *El profesional de la información*, 2006, septiembre–octubre, v. 15, n. 5, pp. 384–389.

Introducción

EL MOTOR DE BÚSQUEDA *GOOGLE* es reconocido de forma universal como la más completa, potente y eficiente herramienta de recuperación de información en internet. Aunque tanto desde ámbitos académicos como comerciales se han criticado algunas de sus prestaciones, sigue siendo el referente básico para los estudios de contenidos en la Web.

En los últimos años *Google* ha liderado la llamada guerra de los buscadores, una disputa centrada en el volumen de información que recogía cada una de las bases de datos de los principales motores, y en la que a pesar del baile de actores (*Yahoo* heredó *Altavista* y *Alltheweb*, *MSN Search* se posicionó recientemente) *Google* parece haber mantenido la primera posición.

Posición	TLD	Mill. págs.	Posición	TLD	Mill. págs.
1	*com*	15.830	17	*ch*	287
2	*org*	4.840	18	*se*	258
3	*edu*	2.840	19	*br*	253
4	*gov*	1.790	20	*cz*	235
5	*de*	1.560	**21**	***es***	**215**
6	*jp*	1.430	22	*no*	197
7	*net*	1.340	23	*us*	194
8	*uk*	1.240	24	*int*	188
9	*cn*	731	25	*kr*	184
10	*ca*	590	26	*at*	170
11	*fr*	560	27	*be*	169
12	*ru*	488	28	*info*	154
13	*it*	417	29	*dk*	152
14	*nl*	343	30	*tw*	144
15	*au*	333	31	*fi*	143
16	*pl*	290	32	*números*	127
		Total			39.612

Tabla 1. Distribución por dominios de los contenidos de Google (datos en millones de páginas web)

Estos datos cuantitativos pueden ser muy útiles para analizar la evolución de la Web y estudiar su composición de acuerdo con distintos criterios. Ello es objeto de una disciplina llamada cibermetría (**Arroyo**, et al., 2005) cuyas aplicaciones se extienden a varios ámbitos.

Los motores de búsqueda ya han sido objeto de estudios cibermétricos (**Aguillo**, et al., 2005), pero el crecimiento explosivo de la Web y la política comercial agresiva de *Google* hacen aconsejable un seguimiento periódico de los principales parámetros de su base de datos.

Recientemente *Google* ha procedido a cambiar y mejorar notablemente su infraestructura, una iniciativa llamada "BigDaddy" (**Cutis**, 2006) y cuyo impacto merece ser analizado en detalle.

Métodos

Al igual que algunos otros motores, permite la utilización de operadores con los que ejecutar estrategias de búsqueda que filtran los resultados según diferentes criterios. Algunos son explícitos como los delimitadores de la dirección y basta con interrogar con la sintaxis adecuada. Por ejemplo, *site:es* recupera todas las páginas bajo domino *.es*. Otros aparecen implícitamente en la url de respuesta del buscador y allí se pueden editar. En la siguiente url:

http://www.google.com/search?lr=lang_ca&q=site%3Acat&hl=en&ie=UTF–8&oe=UTF–8

el valor de *lr* es *lang_ca*, que recupera las páginas escritas en idioma catalán.

En el presente estudio se han utilizado varios delimitadores para conseguir tanto información de la cobertura global de la base de datos como su composición de acuerdo con diferentes criterios: países, dominios e idiomas.

«Se han utilizado unos scripts de desarrollo propio basados en el API de Google que permiten la interrogación automática para un volumen medio alto de peticiones»

Se han utilizado unos scripts de desarrollo propio basados en el *API* de *Google* para la interrogación automática de un volumen medio alto de peticiones. De cada resultado se extrae el valor numérico proporcionado por *Google*, repitiendo el proceso simultánea o consecutivamente y seleccionando el valor máximo obtenido de cada par de consultas. Esto se realiza para minimizar el comportamiento irregular del buscador debido tanto a la saturación o contenido desigual de sus *data centers* como al comportamiento aberrante del algoritmo de extracción numérica, que como bien es sabido sólo proporciona resultados aproximados y redondeados (**Bar–Illan**, 2005).

Los dominios interrogados son todos los disponibles en la lista de la *ISO 3166*, más los 3 específicos estadounidenses (*.edu*, *.gov*, *.mil*), los 13 dominios genéricos internacionales (incluyendo *.eu* y *.cat*) y los números ip del 0 al 255. Los países e idiomas han sido extraídos de las listas proporcionadas en la ayuda del servicio *API*.

La recogida de datos se realizó desde el *Cindoc* durante la segunda y tercera semana de mayo de 2006.

http://www.iso.org/iso/en/prods–services/iso3166ma/02iso–3166–code–lists/list–en1.html

http://www.google.com/apis/reference.html

Resultados

Se han obtenido valores diferentes de cero para 253 dominios de

Figura 1. Montaje del tamaño publicado por Google con datos del Internet Archive (http://www.archive.org)

alto nivel (237 *cTLD* + 13 *gTLD* + *.edu* + *.gov* + *.mil*) y 256 sufijos ip[1].

El resultado total es de 39,6 mil millones de páginas (tabla 1) que es una cifra muy superior a los últimos datos publicados por *Google* en 2005 (8 mil millones) o a ciertas estimaciones (**Gulli**; **Signorini**, 2005). La cifra real puede ser inferior ya que habría que tener en cuenta el alto número de enlaces rotos que presenta últimamente *Google* (5–10%) y el redondeo de los datos, que podría suponer entre un 3–5% de error adicional. La existencia de páginas duplicadas no se considera, aunque es un hecho constatado, pues se trata de direcciones distintas. En todo caso el motor seguiría sin indizar más del 50% de la web pública teniendo en cuenta el bajo solapamiento que todavía se detecta entre motores (**Bar–Yossef**; **Gurevich**, 2006).

El dominio *.com* representa por sí solo el 40% del total, mientras

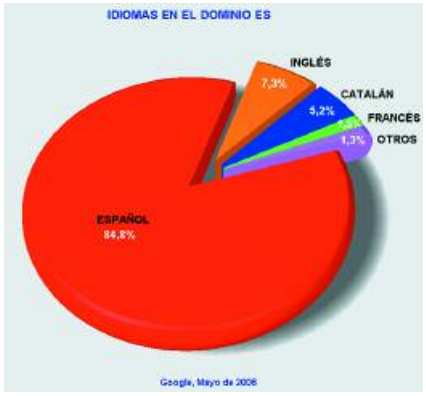

Figura 4. Reparto de contenidos por idiomas en el dominio .es

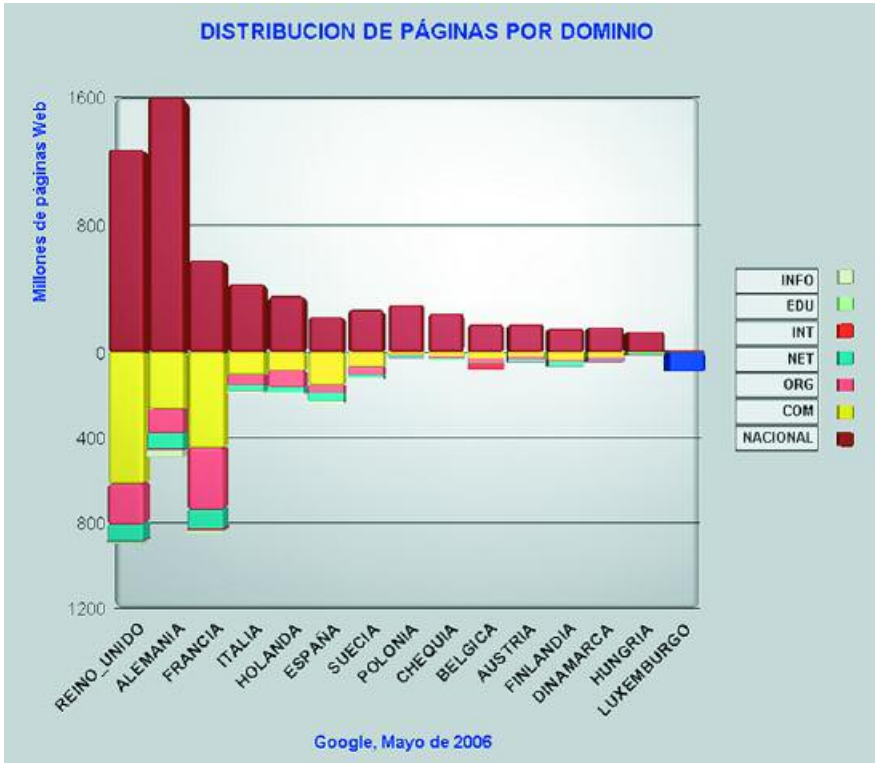

Figura 2. Distribución de páginas por país y dominio (nacional e internacionales).

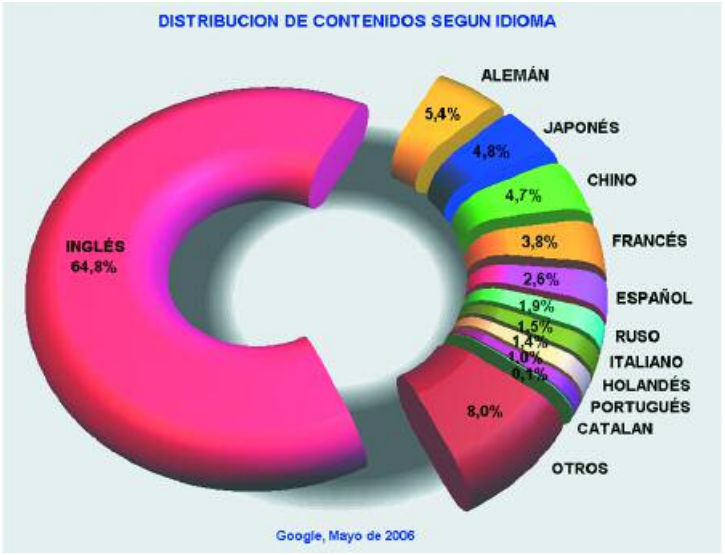

Figura 3. Distribución de contenidos en Google según idioma

que entre los 8 primeros de la lista suman el 78% de los contenidos en *Google*. Ello podría indicar indirectamente la existencia de sesgos en la cobertura (**Vaughan**; **Thelwall**, 2004), especialmente de las regiones en vías de desarrollo.

Sin embargo, la cifra proporcionada para cada *cTLD* no representa necesariamente el volumen total de páginas web de un país, pues en algunos casos la contribución relativa de los dominios internacionales es muy importante. En el presente análisis hemos ignorado algunos dominios nacionales que, como los de ciertas islas del Pacífico e Índico (*.cx*, *.tv*, *.ws*, *.to*, *.cc*, *.fm*) corresponden en realidad a servidores de otros países, siendo el caso más destacado el de las Islas Niue, cuyo sufijo es registrado mayoritariamente por suecos (*.nu* significa ¡ahora! en sueco).

Los resultados (figura 2) muestran que Luxemburgo (85,7%), Francia (59,6%) y España (52,1%) tienen más páginas publicadas en dominios externos que en el propio. Entre los países de la Unión Europea, sólo 5 (todos ellos de nuevo ingreso) tienen más del 90% de sus páginas en su dominio nacional. Distintas razones de tipo político, económico y administrativo explican esta situación, que cabe esperar se acentué con la disponibilidad de nuevos *gTLD* (*.eu*, *.name*, *.cat*, etc.). Es significativo destacar el número no despreciable de páginas

infonomia
LA RED DE INNOVADORES

if... no puede faltar en tu biblioteca

LA REVISTA *if...*

La revista mensual independiente de referencia sobre innovación. Llega a los profesionales y directivos más inquietos de las empresas más innovadoras.

Subscripción corporativa: www.infonomia.com/corp
Subscripción individual: www.infonomia.com/revistaif

Publicítate en *if...* llamando al: 93 224 01 50

Av. Icària 205-207, 2°1ª - 08005, Barcelona - Tel.+34 93 224 01 50 - www.infonomia.com - info@infonomia.com

bajo dominio *.edu*, habitualmente destinado a instituciones de enseñanza universitaria en EUA y que se adjudican universidades de otros países. En los países latinoamericanos el patrón es similar, aunque no tan marcado si exceptuamos la situación de Puerto Rico, donde la gran mayoría de los dominios universitarios son *.edu*.

«Luxemburgo, Francia y España tienen más páginas publicadas en dominios externos que en el propio»

El estudio del idioma presupone que *Google* identifica correctamente de forma automática todos y cada uno de ellos, lo que está lejos de ser cierto. Páginas con algunos párrafos en inglés son habitualmente asignadas íntegramente a esta lengua, lo que genera evidentes sesgos. Además el vasco o gallego no están recogidos explícitamente (aunque *Google* parece incluirlos en español).

Los resultados, obtenidos en mayo de 2006 (figura 3) parecen no corroborar ni una marcada tendencia a la baja del inglés, que sigue representando casi dos tercios del total, ni una creciente importancia del español que no sólo se sigue manteniendo por debajo del 3% sino que ocupa una inesperada sexta posición, detrás de alemán, japonés, chino y francés. Todo ello a pesar de que el número de internautas hispanohablantes representa entre el 7–9% del total.

http://www.internetworldstats.com/stats7.htm

La tabla 2 muestra los dominios con mayor número de páginas en español y catalán, más el nuevo dominio *.cat*. Los porcentajes se han redondeado con el fin de evitar resultados superiores al 100% debido a los errores de aproximación de los datos de *Google* (no al solapamiento, pues sólo se asigna un idioma por página).

Dominio	Inglés %	Español %	Catalán %
.com	77,83	2,41	0,049
.es	7,27	84,81	5,161
.org	84,30	1,97	0,173
.mx	5,04	94,84	0,001
.ar	5,59	94,33	0,001
.net	57,54	3,90	0,372
.cl	6,45	92,87	0
.ve	2,79	96,44	0
.co	8,03	91,62	0
.pe	3,76	95,56	0
.de	8,65	0,68	0,003
.gov	99,12	0,47	0
.edu	96,13	0,24	0,028
.uy	4,81	93,96	0
.int	47,13	3,47	0
.info	50,71	3,43	0,191
.cu	7,91	90,02	0,002
.cr	8,95	89,96	0
.ni	4,11	91,12	0
.cat	3,06	4,66	91,613

Tabla 2. Distribución porcentual de los idiomas españoles en dominios seleccionados

«El estudio del idioma presupone que *Google* identifica correctamente de forma automática todos y cada uno de ellos, lo que está lejos de ser cierto»

Como cabría suponer, el español es el idioma principal en los dominios de países hispanohablantes y tiene una representación entre el 2–4% en los internacionales, a excepción del mayoritariamente europeo (la mayor parte de *.int* es en realidad *.eu.int*, de la Comisión Europea, actualmente migrando a *.eu*). Sin embargo, llama la atención el poco peso porcentual (aun-

que no en volumen de páginas) del español en los dominios *.gov* y *.edu*, lo que delata una ausencia de bilingüismo tanto en el gobierno federal como en las universidades estadounidenses.

En España (figura 4), el dominio *.es* está publicado mayoritariamente por el español, seguido a mucha distancia por inglés y catalán. Este resultado contradice notablemente el obtenido por **Baeza–Yates**, et al. (2005) utilizando un robot que exploraba también páginas bajo dominios internacionales. En dicho trabajo el inglés supone un 30,8% del total, el catalán el 8,2% y el francés el 5,9%, lo que indicaría que en los servidores registrados en nuestro país, aunque bajo otros dominios, el uso del inglés (sobre todo) y el francés se incrementa notablemente (empresas multinacionales, organismos internacionales, etc.).

Discusión y conclusiones

Los datos han sido obtenidos exclusivamente del motor *Google* cuyos métodos de recogida de información y funcionamiento general son secretos comerciales, pero que por diversos estudios empíricos realizados presentan diversas irregularidades y sesgos evidentes. Aún así, es el principal intermediario en la recuperación de información y la sede web más visitada del mundo lo que significa que los datos que proporciona son un referente obligado en la descripción de los contenidos de la Red.

El incremento sustancial en el tamaño de la base de datos parece corresponder a una indización más exhaustiva de la llamada internet invisible o profunda. La incorporación de registros de catálogos de bibliotecas y de bases de datos bibliográficas o alfanuméricas explicaría el salto tanto cuantitativo como cualitativo en los resultados. Sin embargo, la cobertura de la web visible sigue siendo incompleta.

Los datos muestran un gran descontrol en el sistema de dominios, que han perdido gran parte de su carga informativa y que a la luz de los nuevos asignados no hará sino empeorar. Sería deseable la promoción de dominios propios como *.es* o *.cat* o en su defecto la utilización de *.eu* en las condiciones oportunas.

Sin embargo, la conclusión más importante tiene que ver con el estancamiento (en términos porcentuales) que se observa respecto a la contribución de los recursos en español, que eventualmente podría abrir paso a situaciones de colonialismo cultural (fundamentalmente desde el inglés). La única manera de evitarlo es incrementar significativamente la publicación de recursos en nuestro idioma y preocuparse de que dichos contenidos estén indizados y bien posicionados en *Google*.

Notas

1. *cTLD = country top level domain*, *gTLD = generic top level domain*

Bibliografía

Aguillo, I. F.; **Arroyo, N.**; **Pareja, V.**; **Ortega, J. L.**; **Prieto, J. A.** «Análisis cibermétrico de los principales motores de búsqueda». En: *9as. Jornadas españolas de documentación*, *Fesabid 2005 Infogestión*, 2005, pp. 255–272.

Arroyo, N.; **Ortega, J. L.**; **Pareja, V.**; **Prieto, J. A.**; **Aguillo, I. F.** «Cibermetría. Estado de la cuestión». En: *9as. Jornadas españolas de documentación*, *Fesabid 2005 Infogestión*, 2005, pp. 273–289.

Baeza–Yates, R.; **Castillo, C.**; **López, V.** «Characteristics of the web of Spain». En: *Cybermetrics*, 2005, v. 9, n. 1, paper 3. *http://www.cindoc.csic.es/cybermetrics/articles/v9i1p3.html*

Bar–Ilan, J. «Expectations versus reality–search engine features needed for web research at mid 2005». En: *Cybermetrics*, 2005, v. 9, n. 1, paper 2. *http://www.cindoc.csic.es/cybermetrics/articles/v9i1p2.html*

Bar–Yossef, Z.; **Gurevich, M.** «Random sampling from a search engine's index». En: *WWW 2006 Conference*, 2006. *http://www2006.org/programme/files/pdf/3047.pdf*

Cutts, M. *Feedback on Bigdaddy data center* (entrada del 4 de enero de 2006). Bitácora de **Matt Cutts**: *Gadgets, Google, and SEO*. Consultado en: 10–05–06 *http://www.mattcutts.com/blog/bigdaddy/*

Gulli, A.; **Signorini, A.** «The indexable web is more than 11.5 billion pages». En: *Proceedings of 14th International world wide web conference*, 2005, pp. 902–903. *http://www.di.unipi.it/~gulli/papers/f692_gulli_signorini.pdf*

Vaughan, L.; **Thelwall, M.** «Search engine coverage bias: evidence and possible causes». En: *Information processing & management*, 2004, v. 40, n. 4, pp. 693–707.

Isidro F. Aguillo, ***José L. Ortega***, ***Begoña Granadino***, *Grupo de Investigación en Cibermetría, Cindoc–CSIC, Joaquín Costa 22, 28002 Madrid. {isidro,jortega,bgranadino}@cindoc.csic.es*