El libro de mil millones de páginas (La ecología lingüística de la Web)

Posted on 1 de septiembre de 2000 by José Antonio Millán

Que otros vean la Internet como un gran despliegue de anatomías. Yo voy a considerarla como un vasto océano de palabras. ¿Cuántas? En estos momentos, y según Inktomi, en la red podría haber cerca de cinco millones de sitios *web,* que totalizarían más de 1.000 millones de páginas, en todas las lenguas del mundo (incluyendo algunas lenguas muertas y otras inexistentes…). Pero sobre todo están en inglés: el 86%. Hay también cerca de 250 millones de imágenes, no todas ellas pornográficas.

Para tener una idea de esta magnitud supongamos que alguien quisiera leer todo el conjunto de la Web a fecha de hoy (preciso esto porque los contenidos de las páginas están variando cada día, lo que hace que la masa total de páginas *diferentes* en períodos más amplios aumente aún más). Si dedicara a ello una jornada laboral normal, pero sin festivos ni vacaciones, tardaría más de veinticinco mil años… Pero esto no es todo: los contenidos que se vuelcan en los grupos de discusión (*news groups*) puede perfectamente *cuadruplicar* el contenido de la Web. Y por último, el conjunto del correo electrónico que circula por el mundo supera con mucho las cifras anteriores.

A lo largo de la historia humana se han generado incontables sartas de discurso, y muchas de ellas se han remansado en tablillas, muros, papiros y libros. El saber total de la Antigüedad –al menos en su forma final, custodiado en la Biblioteca de Alejandría– se ha calculado en 0,8 terabytes (un terabyte es aproximadamente un millón de megabytes, o *megas*; un mega es algo menos del contenido de un disquete). Los veinte millones de libros de la Biblioteca del Congreso de Estados Unidos ocuparían (sin contar sus ilustraciones) 20 terabytes. La totalidad de un corte *actual* de la Web daría 7,5 terabytes de texto. Sí: la Red es ahora una inmensa biblioteca o, más bien –dada la interconexión que hay entre sus obras– un gigantesco libro. Lo realmente nuevo y asombroso de la situación actual no es sólo la cantidad, sino que esta masa gigantesca de texto sea directa e inmediatamente accesible.

Los esfuerzos por conocerla y navegarla han producido un hecho cualitativamente nuevo: la mediación de sistemas automáticos para la comunicación entre las personas que publican contenidos en la red, y quienes los buscan. Esta mediación, que tiene una gran base lingüística, va a ser el tema de estas páginas.

¿Como conocemos que algo está en la Red? Porque alguien nos lo dice (por ejemplo, con un *email:* «echa un ojo a esto, que te interesará»); porque estamos en una *lista de discusión* sobre un tema y allí nos recomiendan unas páginas; o porque hemos acudido a un sitio *web* que reúne enlaces según cierto criterio. Estos mecanismos de filtrado colectivo –con todo y ser apasionantes– no nos ocuparán ahora. Nos centraremos más bien en los sistemas de mediación automática…

Un ejército de arañas surca la Red. Son mecanismos virtuales al servicio de los buscadores que, siguiendo los dictados de su programación, leen cada página que encuentran en su camino, y mandan las palabras que contienen (junto con información sobre dónde están) a enormes bases de datos. Acabadas de devorar todas las páginas de un sitio, seguirán los enlaces que encuentren hasta llegar a otro, y comenzar de nuevo. Gracias a ellas, los buscadores (de los que hay unos cuatrocientos, aunque sólo una veintena de importancia) pueden responder a las preguntas de los usuarios: «páginas donde esté la palabra *araña* y la palabra *clavo*». *Altavista,* por ejemplo, funciona así, y gracias a su acción contenidos que habrían permanecido ocultos a los ojos de cualquiera se pueden exhumar fácilmente.

Fijémonos que estamos en el dominio de la explotación del interior de los textos: el acceso a los documentos ha venido estando guiado primero por su descripción en archivos y bibliografías, y luego (una vez conseguidos) por las balizas textuales que el autor o el editor fijara (títulos de obras, de capítulos y apartados, índices), pero aquí estamos en el acceso a la palabra, a cualquier palabra del interior.

Los buscadores que indizan el contenido de las páginas nos sitúan frente a la Web –aunque matizaremos esto– como el sabio dotado de un volumen de *Concordancias* frente a las Sagradas Escrituras: en el dominio pleno.

Pero… ningún buscador indiza más allá del 50% de la Red, afirma *Search Engine Watch,* y en estos momentos es difícil saber qué proporción de Web abarca la suma de todos los buscadores (o, en otras palabras, qué porción de la Web es opaca a cualquier búsqueda). En febrero de 1999, un estudio de la revista *Nature* (recogido en *WWWMetrics*) calculaba que entre todos los buscadores no se cubría más que el 42% de la Red. La situación habría mejorado, pero en cualquier caso parece mantenerse la tendencia manifestada entonces (se indizan sobre todo sitios de EE.UU., y en especial los más visitados, y los comerciales más que los de educación). Como consecuencia, o las lenguas minoritarias en la Red se dotan de herramientas propias, o quedarán en una zona de sombra…

Hay sitios dedicados a monitorizar las demandas que los usuarios dirigen a los grandes buscadores (recordemos: ¡todo en la Red es transparente!), y produce una extraña impresión

asistir al rosario de peticiones que se van desgranando hora a hora: se puede hacer la experiencia en *Wordtracker*. Esta misma empresa ha calculado cuántas consultas a los buscadores se hacen diariamente: el asombroso resultado son 250 millones. A partir de ellas *Wordtracker* crea una base de datos de 30 millones de términos (las palabras y frases por las que busca la gente), que se renueva cada dos horas. Luego hablaremos de la importancia comercial que tiene conocer estas búsquedas.

¿Y qué es lo que persigue la gente en la Red? Hay una demanda clásica, mantenida constantemente desde que se empezó a estudiar la exploración de la Web: *sex* (y podemos sentir aquí ecos del Arcipreste: «El hombre por dos cosas se mueve, la primera…»). Según *Mall-Net, sex* es la palabra que lanzan más del 5% de las personas que se asoman a un buscador. Durante años ha sido la más pedida: sólo recientemente ha sido superada por *MP3* (el famoso formato de compresión en el que circula la música).

Algunas palabras o frases adquieren notoriedad durante unas horas o días, y luego desaparecen de pronto. *Mall-Net* las registra y suele comentarlas. Por ejemplo: el 15 de mayo pasado hubo un pico de demandas sobre *car security*. La causa más probable fue que el primer ministro de Cachemira (junto a cinco personas más) murió por una bomba puesta en su coche. Las demandas subieron una hora después de que un despacho de Reuters diera la noticia.

Aun con la ayuda de las grandes bases de datos de los buscadores, localizar algo en la Red no es tarea fácil; exige habilidades detectivescas de un tipo muy especial (y que antes estaban confinadas sólo a profesionales de la documentación). Hay que saber qué expresión es más probable que aparezca relacionada con el material que se persigue, huir de los términos que se pueden emplear en más de un campo, saber combinar varias palabras clave mediante operadores booleanos –Y, O, NO– para refinar la búsqueda (*Lutero* Y NO *King*). Todo esto ya se empieza a enseñar en los colegios (avanzados).

Mientras tanto, ya hay *software* que permite que las búsquedas se abran morfológicamente (preguntar por *conducir* y acceder también a *conduje* y a *conducción)* y semánticamente y por variantes de lengua (de *conducir* llegar a *manejar* y *guiar*). El objetivo último es que se pueda llegar a consultar los buscadores usando directamente la lengua natural, sin tener que reducirla a fórmulas lógicas. Al fin y al cabo, la gente ya tiende a formular búsquedas mediante pequeñas expresiones (el 67,5% de las personas preguntan mediante fórmulas de dos o más palabras, según *Search Engine Watch*).

El siguiente paso es romper los límites de la lengua. Hay buscadores que suplementan la demanda del usuario con traducciones (y al preguntar por *conducir* se buscaría

simultáneamente *conduire* y *drive*), de modo que la consulta llevará a sitios en varias lenguas. Un *software* de traducción, incorporado ya en muchos casos al mismo buscador, permitirá que el consultante poco políglota se entere (aproximadamente) del contenido de esos sitios... Y por último, ya hay programas que hacen resúmenes aceptables. ¿Buscar algo y encontrarlo en cualquier lengua? Cada vez es más posible...

A propósito (y aunque no es el objeto directo de este artículo): del panorama descrito se deduce en seguida que las tecnologías de la lengua van a tener una gran importancia económica y estratégica. La pregunta «¿a quién pertenece el español?» tenía hasta ahora la respuesta retórica (pero real): «¡al pueblo que lo habla!». Pero su prolongación digital, «¿quién va a poser las tecnologías lingüísticas que permitirán a los sistemas automáticos usar el español?», puede tener respuestas desagradables para nuestra economía y soberanía cultural (y la de los demás países hispanohablantes, por cierto…).

El buscador se convierte –así– en el intermediario, en el filtro por el que los contenidos de las *webs* llegan al público. Imaginemos que la única vía de acceso a todas las publicaciones mundiales fueran los ficheros de una decena de bibliotecas. En la Web estamos así, con la diferencia de que en los buscadores no ha habido bibliotecarios que apliquen sus saberes clasificadores: sus programas actúan ciegamente, analizando el código de la página, y clasificando los sitios según criterios formales. Este último aspecto es básico: cuando puede haber cientos o miles de sitios que responden a una determinada búsqueda, figurar en los primeros puestos de la lista de respuestas de un buscador puede ser la diferencia entre existir o no. No extrañará, por tanto, que haya quien lo intente por todos los medios.

¿Cómo extraen los buscadores la información sobre qué contienen las páginas? Los sitios *web* tienen zonas reservadas a la comunicación con los robots: hay anotaciones internas (llamadas *metatags*) que pueden resumir el contenido del sitio. Normalmente los buscadores dan más peso a una palabra clave con la que el gestor del sitio ha descrito su contenido, que a una palabra que han cogido del cuerpo del texto. También valoran las palabras que aparecen en un título, o que están más veces en un documento. Para figurar en los primeros puestos de un buscador caben distintas estrategias legítimas, pero la importancia económica de la clasificación ha generado también el *spamdexing*.

Si spamming era la práctica de enviar correos electrónicos no solicitados, aquí el término se alía con *indexing* («indización», que es lo que hacen los buscadores). Se considera *spamdexing* los procedimientos para provocar artificialmente que un sitio *web* aparezca en una posición destacada en un buscador (o se liste varias veces), cuando el consultante pregunta por determinadas palabras clave. Se consigue metiendo en los *metatags* repeticiones («sex sex sex fuck fuck fuck…», como hacen los sitios *porno*), pero también

colocando en el cuerpo principal de la página semejantes acumulaciones. Para que no perturben al lector, se incluyen frecuentemente en el mismo color que el fondo de la página, con lo que resultan invisibles… aunque no para el robot. Si prolongamos la imagen de la Web como biblioteca, esta práctica sería equivalente a someter un libro llamado *Galdós Galdós Galdós*, con un índice que fuera: «Con Galdós / Sobre Galdós / Por Galdós…», y comenzara: «Esta obra trata sobre Galdós y la obra de Galdós». Si cayera en manos de un catalogador poco profesional lo clasificaría en… «Galdós», y si tuviera que ordenar los libros por su importancia para el tema lo colocaría muy arriba…

Esta estrategia podríamos llamarla «de repetición» o «del niño de cinco años» (en honor a su sistema de obtención de cosas: «¡Quiero un pokemon, un pokemon, lo que quiero es un pokemon, un pokemon!»). Pero existen otras. Está, por ejemplo, la estrategia «del rape», en honor a este pez (y a los de otras especies) que fingen formas de pececillos en sus protuberancias, para atraer a depredadores… y devorarlos. En esta estrategia un sitio se mimetiza parcialmente con otro, cuyo tráfico quiere conseguir. Esto se realiza incluyendo el eslogan o la marca del competidor (de forma patente o en *metatags)*: cuando los usuarios preguntan por la competencia, el buscador devuelve el sitio sin escrúpulos (incluso por encima del legítimo, si las estrategias que éste tiene para el diálogo con los buscadores son peores…). Continuando con la imagen de la biblioteca, el equivalente sería que el libro sobreabundante en Galdós que hemos entregado, en realidad contuviera en su interior una novela de… Clarín.

Y hay un caso extremo, que llamaremos «de la hormiga esclavizadora», por ciertas especies que roban huevos a las vecinas y cuando nacen las crías las ponen a trabajar a su servicio. Éste es el caso de quienes copian una página de un sitio y la meten en su propio servidor. El buscador indiza la página ajena con todas sus claves legítimas, y cuando el usuario pregunta por ellas el buscador le devuelve tanto la dirección de la página verdadera como la de la falsa. El usuario que hace clic en esta última llega a la página requerida… en un servidor extraño, donde rápidamente se le reenvía a un nuevo, y no buscado, contenido. En un caso reciente, si uno buscaba «libros medievales para niños» el buscador devolvía, entre otros, dos enlaces a *Gloriana's Book Store* con la siguiente descripción, «libros medievales para jóvenes lectores». Uno de ellos era legítimo, pero el otro conducía al sitio porno *http://*www.taboosisters.com. En la metáfora libresca, sería equivalente a arrancar una portada de *El pequeño príncipe*, y pegarla en un ejemplar del *Decamerón*, para colocarlo en el mostrador de la librería… Los tribunales norteamericanos abundan últimamente en demandas contra todo tipo de *rapes* y *hormigas.*

Hemos mencionado antes las marcas y los eslóganes. Estos fragmentos de lengua (nombres propios y frases) se están convirtiendo en un patrimonio aún más valioso en un medio, como

la Web, en que la textualidad y la literalidad son claves en la localización. El grupo activista Irational de Net-Art (o arte basado en la Red) ha satirizado la tendencia a la hiperprotección en su página *web,* que lleva el siguiente texto:

«Por favor, tenga cuidado con su lenguaje. Las siguientes frases son marcas registradas, y son propiedad de sus propietarios registrados.
Recuerde: el lenguaje no es libre TM .
Todo está conectado TM .
Mire, oiga y sienta la diferencia TM .
Soluciones para un planeta pequeño TM .»

Y así a lo largo de unas decenas de frases (normalmente banales) que tienen propietario. Naturalmente, la vuelta de tuerca es el (¿pseudo?) registro de la frase «Recuerde: el lenguaje no es libre TM ».

Pero también hay programas residentes en ordenadores personales que apelan al análisis de palabras clave. Son los programas de censura de la Red (*censorware),* también llamados programas-niñera (*nanny software)*, que tienen por objeto vigilar la salida a Internet de los menores de edad. El *censorware* tiene una lista de sitios a los que no permite acceder. Para localizarlos, las arañas de las *ciberniñeras* (que tienen nombres tan expresivos como *MudCrawler,* «la que Repta por el Cieno») surcan el ciberespacio analizando las palabras que contienen los sitios, para irlos colocando en su lista negra.

Pero en este análisis –como en todo análisis exclusivamente formal– acechan los peligros de la homonimia y la polisemia, por no hablar de los usos legítimos de descripciones anatómicas. Es conocido el caso del *censorware* que impedía el acceso a un sitio sobre *cáncer de pecho* y a recetas de *pechugas de pollo*, porque excluía sistemáticamente la clave *pecho* (en el original *brest cancer* y *chicken breast*). Contra los abusos de *cibernannies* como *CYBERsitter* y otras se ha levantado el sitio *Peacefire*, donde se pueden encontrar detalles sobre la actuación del *censorware*.

Otra misión de los programas-niñera es monitorizar las conversaciones y comunicaciones con el exterior del ordenador, ya sea mediante correo electrónico, en *chats*, etc. Por ejemplo: el *software* puede impedir que se dé el teléfono o las señas de la casa, vetar el empleo de determinadas palabras, etc.

Y más allá de la censura doméstica está la institucional y estatal. Desde hace muchos años determinados sitios y servicios de alojamiento de *webs* escudriñan los contenidos de las páginas que contienen, en busca de contenidos ofensivos. Pero quizás el proyecto más

grande de espionaje automático público sea Echelon, que comparten Estados Unidos, el Reino Unido, Canadá, Australia y Nueva Zelanda. Se supone (porque es un proyecto secreto) que se dedica a espiar todas las comunicaciones, incluyendo correo electrónico, teléfono y fax. De tener como objetivo en el pasado las acciones de las potencias enemigas ha pasado al espionaje económico y comercial (mucho más rentable). El tema llegó al Parlamento Europeo en febrero de 2000, pero varios meses antes la comunidad usuaria de la Red ya había detectado esta extensa acción –se supone que Echelon controla dos millones de mensajes al día– y, reaccionaba intentando crear *ruido*. Muchos mensajes que circulaban en el verano del 99 incluían una coda que decía: «drug kill assassinate spy plutonium president security bomb hijack heroin». La acción culminó el 21 de octubre, con una llamada masiva a hacer estos envíos. Supongo que los ordenadores de Echelon se vieron momentáneamente sobresaltados… hasta que metieron una rutina para descartar esa sarta.

Dos cuestiones se suscitan aquí: una es la transparencia de la comunicación por la Red (cualquier correo electrónico puede ser leído, alterado, reenviado; cualquier conversación en un chat puede ser interceptada…), y la segunda es el grado de sutileza de los análisis automáticos. Los buscadores no van más allá de una identificación formal, pero podemos pensar que el software de espionaje debe ser más avanzado, y podrá «entender» los textos en cierta medida… Algunas comunicaciones por correo, y determinadas páginas web intentan burlar los sistemas automáticos de detección mediante pequeños cambios: *f*ck*, *h*j* d* p*t*.* Pero hoy en día los sistemas basados en reconocimiento de patrones y lógicas difusas sortean muy bien estos pequeños maquillajes.

Y otra cuestión: ¿realmente los terroristas y narcotraficantes son tan explícitos en sus comunicaciones? Podemos asegurar que no. Pero sin embargo hay otra modalidad de comunicación que debe ser muy clara: aquella que se ofrece para destinatarios desconocidos. Por ejemplo, los grupos que ofrecen recetas para bombas caseras. Precisamente una demanda masiva de «yellow phosphoric nitric» en los buscadores, que observó *Mall-Net* a mediados del pasado mayo, se puede atribuir a un intento del FBI por detectar estos sitios.

Hoy en día los textos digitales forman un *continuum* cuyo ecosistema se distribuye entre la red y los ordenadores personales e institucionales. Como ejemplo, tenemos lo que ocurre en un terreno menos espinoso que los que hasta ahora hemos recorrido: los trabajos estudiantiles. Desde que existe la WWW, es muy sencillo acceder a cualquier página web y coger información que luego se firma con el propio nombre y se entrega al profesor (también se hace esto y se entrega al director de la revista o el periódico: no infamemos sólo a los pobres estudiantes…). A ello se une la existencia de sitios destinados a proporcionar trabajos de curso completos, sobre los más diversos temas (en español tenemos uno,

significativamente llamado *El rincón del vago*). De hecho, fuentes académicas reconocen que la copia de trabajos ha aumentado sin parar desde 1990, y el *cybercheating* (cibercopieteo) es directamente responsable de ello.

¿Cómo se puede saber si un trabajo es una copia, total o parcial? Un estudiante de Berkeley creó un servicio, *Plagiarism*, que permite a los profesores indagar sobre la paternidad de los trabajos que les entregan. El creador de este servicio afirma –un tanto cínicamente– que ha equilibrado la balanza, y facilita a los educadores el mismo poder que los alumnos ya tenían gracias a Internet. Sea cual fuere la valoración moral que merezcan una y otra práctica, funciona. El primer test se hizo con 300 estudiantes de neurobiología, y aunque se advirtió a los alumnos que sus trabajos serían rastreados, se encontró un 15% de copias. *Plagiarism* compara los textos que se le someten con distintos rastreos de la Web a cargo de veinte buscadores, y con bases de datos de trabajos, que incluyen los que aportan las mismas universidades que contratan el servicio. El otro gran flanco de visibilidad en Internet (y simétricamente, de estrategias de búsqueda) son los nombres de los dominios. Dominio es la sarta de letras que aparece a la derecha del signo @, en la direcciones de correo, y tras *http://* y antes de */* en las direcciones web. Hay dominios llamados «de primer nivel» (*.com*, *.org,* ...) y dominios territoriales, que corresponden a los países (*.es* para España, *.fr* para Francia…). De los primeros hay registrados un total de diez millones, y de los segundos tres millones.

El hecho trágico, e imprevisible cuando se constituyó el sistema de dominios, es su tremenda escasez actual. En un momento en que el espacio de almacenamiento es prácticamente gratuito (por toda la Red se regala espacio para páginas *web*, o para almacenar las fotografías de uno…), el recurso más directamente lingüístico, el nombre del dominio, está prácticamente agotado. En el dominio más universal, *.com,* hace ya muchos meses que todas las combinaciones de cinco letras están cogidas. Pero hay más: un estudio de *Wired News* en abril de 1999 señalaba que de las 25.500 palabras del inglés estándar sólo 1.760 no se habían convertido aún en nombres de dominio. Eso significa que los términos ingleses para «árbol» o «dinosaurio» o «calcetín» o casi cualquier otra cosa ya tienen dueño. Y por supuesto, cientos de miles de nombres propios. Puede que algunos de ellos estén en manos de *ciberocupas* (personas que compran dominios que no usan, para luego revenderlos), pero muchos tienen un ocupante legítimo, y el único problema es que sólo se pueden usar una vez. Naturalmente, la carestía es mayor cuanto más breve es el dominio, y el caso extremo son las siglas: cuando la British Broadcasting Corporation (BBC) quiso un dominio *bbc.com* tuvo que pagar bastante dinero por él, ya que pertenecía, con toda legitimidad, a la empresa estadounidense Boston Business Computing.

El problema con las marcas registradas era el mismo, pero con un matiz: que sus nombres se

podían repetir en distintas clases (por ejemplo, la editorial Taurus y los electrodomésticos Taurus). En la Internet sólo hay una clase… por el momento, porque hay planes de introducir *.sex, .firm, .shop, .web, .arts, .info*…, que aumentarían el acervo de dominios.

¿Qué hacer cuando uno encuentra que un dominio que necesita para un proyecto está *cogido*? Aguzar el ingenio. Cuando el artista y creador multimedia afincado en Francia Laurent Sauerwein quiso registrar *pixel.com* lo encontró ya ocupado, de modo que se inventó un inexistente femenino francés *pixelle.com* (que suena igual). El especialista en videojuegos Javier Candeira no pudo registrar *interactivo*, y optó paronomásicamente por *hiperactivo*…

Cualquier accidente, cualquier circunstancia se aprovecha rápidamente para aumentar las posibilidades de nombrar dominios. Por ejemplo, los dominios territoriales (siglas de los nombres de países) pueden dar lugar a combinaciones interesantes en diversas lenguas. Un proveedor español registró en Alemania *–.de– pagina.de*, y comercia con ese servicio (al que se puede alquilar, por ejemplo, la dirección de web). Pero el caso más reciente y llamativo ha sido el del país polinesio de Tuvalu, que ha vendido su dominio *.tv* por 50 millones de dólares anuales en los próximos doce años (imaginemos el interés de ser propietario de *guia.tv* y similares…). Paradójicamente, el *nombre* del país se ha convertido en su recurso número uno.

Y es que el dominio es realmente importante. Muchas personas abordan sus búsquedas desde él. Por ejemplo, alguien que quiera comprar una entrada se sentirá tentado de empezar por *http://www.entradas.com*. Los buscadores también tienen tendencia a dar una buena posición a las direcciones web que coinciden con la palabra buscada. También en la dirección de las páginas se asiste a prácticas de *spamdexing.* Véase este ejemplo detectado por Inktomi (*tax* es «impuesto» en inglés):

http://www.tax.taxadvice.taxation.irs. taxservices.taxrepresentation. taxpayerhelp.internalrevenueservice. audit.taxes.com

Y a la inversa: una de las cosas que hacen los programas de *censorware* es vetar los sitios que tienen algo sospechoso en su dominio (llegando al extremo de rechazar los que tienen varias X, signo de «contenido para adultos» en inglés). Como muestra del interés económico del tema, existe un servicio *web* llamado *DomainCheck,* que se dedica a analizar el contenido de los proyectos que se le someten, lo contrasta con las palabras clave más pedidas a los buscadores, y con el resultado propone una combinación o amalgama.

La *ciberocupación* no sólo afecta a las marcas comerciales (registrándolas antes que su dueño, como pasó en España con Hipercor). Véase lo ocurrido con el nacimiento del hijo del

primer ministro británico, al que se dio el nombre de Leo Blair. Pocas horas después del anuncio del nombre ya estaban registrados *leoblair.com* y *babyleo.co.uk*, por personas que intentan hacer un negocio con su reventa…

Y por fin otras estrategias tienen que ver con la modificación hacia el insulto. Si quiero molestar a los propietarios de *patata.zu*, registro *patatapodrida.zu*. Las estrategias de defensa entonces se deben extender a terrenos insospechados, por ejemplo: adelantarse a los enemigos. En la reciente campaña norteamericana del gobernador Bush, parte del presupuesto de 60 millones de dólares se ha invertido en comprar sitios como *bushsucks.com, bushblows.com…* (versión libre: «Bush me la sopla»). Quienquiera que entrara en esas señas llegaría al sitio oficial, en vez de a la esperada página de ataque al candidato.

La protección de un nombre debe llegar incluso a sus erratas. La empresa *Typo.Net* las registraba (por ejemplo, *mircosoft.com,* luego vendida a un buscador), para luego redireccionar al descuidado navegante a la página correcta… mientras le colocaba un par de anuncios. Pero una grafía próxima como ésta puede ser utilizada también por un sitio que intente desprestigiar a la marca. Estamos en un terreno en el que, paradójicamente, la agudeza verbal, el juego de palabras y el insulto han adquirido de nuevo la importancia que tuvieron en épocas más orales de la cultura.

Y a todas éstas: ¿qué contiene la Internet? Respuesta: cualquier cosa… Recientemente he leído esta cínica frase: «Alguien dijo que si un millón de monos golpearan una máquina de escribir durante mil años acabarían por escribir las obras de Shakespeare. La Web demuestra que esto no es así». Yo añadiría que muchos otros medios lo demuestran también: ¿cuál es el porcentaje de excelsitud entre los 50.000 libros que se editan anualmente en España, los miles de números de periódicos y revistas, las horas de radio y televisión?

Sea como fuere, hay aspectos que hacen que valga la pena conservar lo que hay en la Internet, y conservar lo más posible. ¿Por qué? Pensemos por un momento: ¿cómo saber qué es lo realmente *valioso* de los contenidos que hoy nos circundan, en cualquier soporte? La respuesta es: ¡vaya usted a saber! El ejercicio de ver qué queda de los autores más vendidos de hace veinte (o cien años) ya es de por sí una lección de humildad. Las bibliotecas resolvieron ya hace años este espinoso problema mediante un sencillo artificio: el depósito legal, que hace que reciban (por lo menos en teoría) un ejemplar de *cada* publicación para su preservación. ¿Por qué no hacer lo mismo con la Web?

Y otro argumento: para los estudios sociológicos, antropológicos, lingüísticos o culturales, cualquier material es aprovechable. Hoy en día estudiamos los *graffiti* de las letrinas de

Pompeya por la información que bindan sobre el latín vulgar: ¡qué no daríamos por poder oír una conversación de la época…! Pues bien: muchos rincones de la Web actual, incluso desdeñables según parámetros culturales, pueden ser una mina para el investigador futuro…

Un tercer argumento para preservar la Web es la inmensa volatilidad de los enlaces. Hace unos años se calculó que la duración media de cualquier enlace era de 44 días, y las cosas no parece que hayan mejorado mucho desde entonces (frente a 4,2 millones de sitios accesibles que examinó Inktomi en febrero de 2000, otros 0,7 millones eran inalcanzables). Es decir: es posible que una información a la que ya accedimos, o que nos han referenciado correctamente, haya desaparecido de la Web, por muerte de su servidor, o sencillamente por traslado.

En la conciencia de todos estos problemas, hace seis años Brewster Kahle puso en marcha el proyecto *The Internet Archive*, con el propósito de preservar las mayores porciones posibles no sólo de la Web, sino también de grupos de discusión o *newsgroups*. Por el momento ya ha almacenado 14,5 terabytes, de los que la mayor parte son páginas *web* conseguidas por el rastreo mediante arañas, aunque también se admitan donaciones. Distintas instituciones y archivos hacen uso del proyecto; por ejemplo: la Smithsonian Institution ha estudiado allí los sitios *web* de elecciones de EE.UU. de 1996, que si no habrían desaparecido para siempre…

Bien: llegados aquí tenemos pleno derecho a imaginarnos la Web como un hormiguero transparente, donde cientos de miles de páginas nacen y mueren cada día, surcado por las arañas (de los buscadores, de las ciberniñeras, de los supragobiernos, de los proyectos de preservación), que devoran y regurgitan ordenadamente su contenido; donde cada segundo tres mil personas arrojan una petición al dios de las palabras, esperando encontrar su camino; en el que mentes aviesas, *okupas* y multinacionales colocan señuelos textuales para confundirles y atraerles… Y al tiempo, no dejemos de ver en ella la biblioteca, el libro gigantesco, que estamos empezando a pensar cómo hojear.

En palabras de Steve Jones, compilador del volumen *Doing Internet Research. Critical Issues and Methods for Examining the Net* (Londres, Sage, 1999): «La Internet es un "tipo distinto de objeto" (si es que constituye, en realidad, un objeto), y estudiarlo requiere un "cambio consciente de foco"».

Ojalá estas páginas hayan contribuido a la tarea de crear un nuevo foco para un nuevo medio.

REFERENCIAS

Altavista < http://www.altavista.com>
CYBERsitter <http://www.cybersitter.com>
DomainCheck
El rincón del vago <http://www.rincondelvago.com>
Inktomi < http://www.inktomi.com/webmap.htm>
Internet Archive
Irational
Mall-Net
MudCrawler
Peacefire
Plagiarism
Search Engine Watch <http://www.searchenginewatch.com>
Typo.Net
Wired News:
Wordtracker:
WWWMetrics: