La World Wide Web nos gusta tanto que tenemos tres

Recomendaciones y advertencias para la construccion de un Archivo de la Web en español

Javier Candeira
javier@candeira.com

Un chiste sobre Internet dice que hay tres copias de la World Wide Web. Una que existe en los buscadores como Google y Yahoo, una segunda en el archivo de Internet disponible en www.archive.org, y la tercera, que es una copia menos fiable pero distribuida que está repartida por servidores de todo el mundo.

Este chiste, como tantos otros, tiene una gruesa capa de verdad. Pese a que nos parezca meramente un repositorio de informacion, la Web es principalmente un mecanismo de transmision por el que accedemos a todo tipo de servicios. Los compendios de informacion son uno de esos servicios, pero también existen sitios de comercio, de juegos,administraciones herramientas editoriales públicas y privadas, buscadores, y archivos.

Estos dos últimos servicios, los buscadores y los archivos, son meta-servicios, que sirven al público y los usuarios manipulando la propia Web, masajeándola, convirtiendo parte de su contenido en valor. Parte de este valor replica simplemente algunas de las funcionalidades de la red, pero otros son nuevos.

Las tres World Wide Webs

		La WWW	***Buscadores***	***Archivos***
Funciones	***Informacion***	Actualizada	Sincrónica	Diacrónica
	Servicios	Todos*	Búsqueda **	Archivo Web **
	Archivo	Opcional, limitado	Limitado	Total***
	Búsqueda	Limitada	Total***	

* Una de las características de la Web es que funciona como meta-protocolo, y sobre ella se están montando todo tipo de servicios, no solo relacionados con el mundo físico (compras, catálogos de bibliotecas, gestiones administrativas) como de acceso a otros servicios online (correo, chat, juegos).
** Las empresas de búsquedas se dediquen tambien a otros negocios, pero un buscador sólo busca.
*** Por su propia definición, no necesariamente por el alcance de sus logros.

Veamos algunos ejemplos:

Un sitio web, un buscador, un archivo

		Elmundo.es	***Google, Yahoo***	***Archive.org***
Servicios	***Informacion***	Actualizada	Sincrónica	Diacrónica
	Archivo	Tres veces al día	Último caché	Exhaustivo
	Búsqueda	Limitada	Exhaustiva	Limitada

Cómo hacer un archivo de la Web: tres pistas

Realizar alianzas tecnológicas (“Zapatero, a tus zapatos”).

En el punto medio esta la virtud, y seria un error pasar del “que inventen ellos” de Larra al constante reinventar la rueda (ejemplificado en el síndrome de “not invented here”, el cual, curiosamente, se puede atribuir en su origen al ejército norteamericano, al igual que Internet).

Al problema de la invención sigue el de las operaciones: un Archivo de la Internet en español debería concentrar nuestros esfuerzos en la parte de “en español”, no en la de Archivo de Internet” con sus problemas de almacenamiento de datos, servidores redundantes, etcétera. Por esta razon sería conveniente contar con un aliado tecnológico para poder centrar nuestros esfuerzos en la parte lingüística.

Archive.org es una entidad sin ánimo de lucro, colaboradora con la Biblioteca del Congreso de los Estados Unidos, y participante en otros proyectos como la Open Library del Gutenberg Project y Yahoo!. Ha demostrado tener el conocimiento tecnológico de saber hacer archivos, y la inteligencia institucional de saber colaborar con entidades gubernamentales, entidades no gubernamentales, y empresas. Es, por tanto, el socio ideal para este proyecto.

Dar valor añadido (recordar la parábola de los talentos)

El primer problema que tendría un tal archivo sería el reconocimiento previo al archivo de que sitios web están en español, incluyendo las partes en español de servicios multilingües, estén en el dominio .com, .edu u .org, y su clasificación dialectal e histórica.

Estas tecnologías lingüísticas no sólo se pueden aplicar en la recogida de datos; también se pueden aplicar sobre el corpus recogido. Los servicios de archivo de la Web en español serían innecesarios si sólo fueran de archivo, puesto que ya hay archivos dedicados a archivar toda la W eb. El seleccionar la Web en español y la mayor exhaustividad son valores añadidos, pero si fueran los únicos sería una gran oportunidad perdida.

Dejar que el público haga (“Si lo construyes, vendrán”)

La Internet y la Web son dos ejemplos señeros de construccion colectiva de conocimiento y servicios. El Archivo de la Internet en Español debería seguir los ejemplos de sus hermanas mayores, conlo que abrir el archivo al público debería ser la primera prioridad del servicio.

Por “abrir al público” se entiende no sólo tener un sitio web abierto a los usuarios, sino además:

a) proveer de APIS de programacion para acceso automatizado (ejemplo: color picker de flickr, mapas de pisos en alquiler/venta con Google)
b) abrir el archivo a investigadores e instituciones ajenas al mismo sin ningún tipo de discriminacion ni cargas.
c) publicar el software lingüístico desarrollado con licencias de software libre para promover su uso y mejora por investigadores.

Cómo no hacer un proyecto de difusion de la lengua y la cultura: dos cuentos cautelares

La RAE y el Corpus

El corpus de la RAE es un buen ejemplo de un buen trabajo hecho a medias.

- Los abundantes textos digitalizados cuyos derechos de autor han prescrito podrían estar ya en línea al completo, disponibles para todos, y no sólo para el corpus.
- La base de datos podría ser accesible mediante APIs públicas para estudiantes e investigadores, fueran independientes o afiliados a instituciones, y no sólo mediante una interfaz web que no permite su uso automatizado.
- Las herramientas desarrolladas por la RAE podrian ser libres para que universidades y empresas las pudieran usar en sus investigaciones y productos lingüísticos.

La Biblioteca Nacional Francesa frente a Google Book Search

La actitud de algunas Bibliotecas Nacionales europeas frente a proyectos como Google Print, que proponen extender el beneficio de la indizacion digital a los libros de papel, sólo puede definirse como “que se fastidie el General, no tomo rancho”.

Abrir la totalidad de los fondos cuyos derechos de autor ya han prescrito a proyectos como Google Book Search, Amazon Search Inside the Book y otros debería ser el proyecto principal de todo bibliotecario, dado que su objetivo, la posibilidad de buscar automáticamente en el contenido de los libros, es algo casi mágico.

En su lugar, la Biblioteca Nacional Francesa propone un sistema centralizado y propio de digitalizacion e indizacion, el cual sólo puede ir a cola de proyectos abiertos y globales como la Open Library (en colaboración con Microsoft, Yahoo y el Internet Archive) o el propio Google Book Search, proyecto emprendido en colaboracion con las bibliotecas de Stanford y Harvard.

El proyecto Gallica, que desde la propia Francia se ha criticado como “quizá no el mejor ejemplo de indexacion abierta”, es una muestra de todos los errores que se pueden cometer:

- frames dentro de frames, páginas no señalables con marcadores
- cada libro escaneado en pdfs de una sóla página
- los libros son páginas escaneadas en calidad ligeramente superior a la calidad fax
- no es texto buscable, sino imágenes.
- copyright (perdón, "los droits d'auteur")restrictivo: toda reproducción no privada que no sea de citas breves requiere de permiso previo de la Biblioteca Nacional de Francia para su uso, incluso en obras cuyos derechos han prescrito.

Este modelo representa un triple freno para la difusión de la cultura francesa. Primero, no realiza bien la tarea que se propone realizar, al generar un servicio de menor calidad. Segundo, ocupa el lugar y el presupuesto de quien pudiera hacerla, ya que la BNF no puede tener dos proyectos con el mismo objetivo. Y tercero, y quizá más importante, impide por medios técnicos y legales que otras partes puedan realizar esa difusion.

Si el enemigo percibido es la difusion de la cultura anglosajona, las armas empleadas en internet deberian ser las mismas que las suyas: apertura, difusión y permisos amplios de uso.

Otros retos

Buscar y encontrar

Hay mucha informacion en la red que no es fácilmente recuperable o accesible, por ejemplo:

- Sitios web gratuitos pero de indscipcion obligatoria (La Vanguardia)
- Sitios web que guardan informacion de sesion en la direccion URL
- Sitios web que guardan textos en informacion pública o disponible en papel, pero que en su formato electronico están publicados en intranets, detrás de firewalls o bajo subscripción.

La materia oscura de la red

Asímismo está toda la "materia oscura de la red", la plasmacion de esa práctica diaria de la lengua en correos electrónicos, en los chats y en los juegos en red multijugador, por poner tres ejemplos.

Muchas de estas comunicaciones on privadas, pero muchas otras se llevan a cabo en listas de correo o canales de chat abiertos al público, y recopilarlas y archvarlas sería de gran ayuda para estudiar la lengua y la cultura de nuestro tiempo.

Herramientas libres para una lengua libre

Otro reto es la liberacion de herramientas, tanto lexicográficas como informáticas, para que estén completamente accesibles para los propios hablantes, usuarios y estudiosos del español. Este acceso debería estar abierto tanto tanto a los ciudadanos particulares como a las insituciones públicas, a las empresas como las universidades y centros de investigacion.

Indizar los libros de papel

Archivar la Web es tratar lo virtual como si fuera físico, y fijarlo y plasmarlo para conservarlo. La accion recíproca de digitalizar los medios físicos para poderlos publicar, transmitir y transformar mecánicamente es la que están acometiendo tanto Google con Google Book Search como Amazon con Search inside the book o Yahoo!, el Proyecto Gutenberg y el Internet Archive junto con muchos otros socios de la Open Content Alliance.

El futuro de la cultura es el futuro del Dominio Público, y la Biblioteca Nacional debe comenzar un proceso de digitalizacion y puesta al público de sus fondos cuyos derechos de autor han prescrito.

Usando el mecanismo no destructivo de la Open Content Alliance, una persona puede digitalizar hasta un libro por hora en una máquina con un costo entre 15.000 y 30.000 euros. Suponiendo cuatro personas a tiempo completo, en dos turnos, durante cinco años de amortización de las máquinas se podrían digitalizar unos 37.000 volúmenes por unos 180.000 euros, con un coste de menos de cinco euros por volumen.

El proceso moderno de edición parte de originales digitales, y el Depósito Legal se podría realizar obligatoriamente en formatos digitales estándar. Un servicio automático puede después liberar al público las ediciones una vez prescrito el plazo de exclusividad de los derechos de explotación.

Y existen ya proyectos de bibliotecas virtuales con gran número de fondos fuera de derechos, como es la Biblioteca Virtual Miguel de Cervantes, que podría cumplir mejor la misión expresada en sus estatutos concediendo amplios permisos de copia y publicación de sus fondos.

Consideraciones finales

Pese a algunos augurios apocalípticos, el español no está en peligro de desaparecer; muy al contrario, es una lengua pujante y viva. Sin embargo algunos de sus usos corren un triple riesgo.

El primer peligro percibido es la ignorancia de nuestras herencia culturales por parte de los propios hablantes de la lengua, Para evitarlo la única solución es una mayor difusion de la cultura, mediante todos los medios técnicos, económicos y legales. Un archivo de la Internet en español que diera los mayores permisos de uso, sin caer en el error del copyright restrictivo, permitiría y fomentaría esta difusion por medio de iniciativas privadas, insititucionales y empresariales.

La lengua es propiedad de sus hablantes, y la cultura es de los pueblos que la comparten. Si una lengua que nadie habla es una lengua muerta, el mismo adjetivo se puede aplicar a una cultura que nadie copia, transmite, y transforma. Haremos bien en proponer para nuestras instituciones culturales regímenes de derecho de autor que fomenten, y no prohíban, la copia, transmision y transformacion de la cultura.

El segundo peligro es la diglosia global. El inglés es de facto la lengua internacional en los ámbitos económico-comercial y científico, algo que no parece reversible a corto plazo. A medio plazo, sin embargo, el desarrollo de las industrias de la lengua con énfasis específico en el español podrían hacer que la traducción automática, el reconocimiento de voz y otros sistemas de procesamiento del lenguaje natural permitieran una mayor compenetracion de la lengua dominante.

Si bien es probable que estas herramientas nunca lleguen al punto de proporcionar un dominio completo de una segunda lengua, su existencia serviría para recortar las diferencias, sobre todo en el ámbito comercial y económico. Aquí el mínimo común denominador de expresiones accesibles mediante traducción automática ligada a las condiciones estándar de contratación podría convertirse en la *lingua franca,* impulsando la investigacion en este tipo de tecnologías.

El tercer riesgo, uno señalado por José Antonio Millán, es que estas industrias de la automatización de la lengua se concentren en unas pocas manos, y "acabemos pagando por usar el español". No abundaré en el argumento, que por otra parte él mismo ha expresado con más lucidez y detalle del que podría expresar yo aquí, pero sí que propondré una solucion: el software libre.

El uso del software libre para la lingüística computacional sirve para garantizar un acceso igual para todos, ciudadanos e instituciones, particulares y empresas, ademas de fomentar la transferencia de tecnologías imprescindible para que todos los hablantes puedan mantener control sobre sus lenguas maternas, para que los hablantes de lenguas minoritarias, y por tanto menos rentables, tengan las mismas posibilidades de recibir atencion que los hablantes de lenguas más populosas.

Sobre el autor:

Javier Candeira es editor y co-fundador de barrapunto.com, el sitio web de noticias y discusión sobre software libre y derechos digitales. Tambien trabaja como asesor de medios de comunicacion sobre edicion digital en entornos web, y es escritor sobre asuntos de cultura y política digitales para medios como Rolling Stone, El Periódico, elmundo.es, la Revista de Occidente y el Boletín de la Institucion Libre de Enseñanza. Por ultimo, ha colaborado en la adaptacion y divulgacion para España de las licencias Creative Commons. Creative Commons es una organizacion sin ánimo de lucro que propone licencias de cesion de derechos como herramientas legales para la difusion y compartición del conocimiento.