BIBLIOTECA DIGITAL GREENSTONE

DEL PAPEL A LA COLECCIÓN

Dr. Michel Loots, Dan Camarzan e Ian H. Witten

Human Info (Bélgica)
Simple Words (Rumania)
Universidad de Waikato (Nueva Zelanda)

Greenstone es un conjunto de programas y aplicaciones de software especialmente diseñados para la creación y difusión de colecciones documentales digitales, el cual le ofrece una nueva forma de organizar la información y publicarla en Internet o CD-ROM. Greenstone ha sido elaborado como parte del proyecto de Biblioteca Digital de Nueva Zelanda de la Universidad de Waikato y actualmente es desarrollado y distribuido en colaboración con la UNESCO y la ONG Human Info. Es un software de código abierto disponible en *http://greenstone.org* bajo los términos y condiciones de la Licencia Pública General de GNU.

Queremos estar seguros que este software trabaje bien para usted.
Por favor comuníquenos cualquier problema que tenga con él a
la siguiente dirección: greenstone@cs.waikato.ac.nz

Greenstone gsdl-2.50 **Marzo de 2004**

Acerca de este manual

El presente documento explica la forma de crear colecciones en CD-ROM a partir de documentos impresos y expone con detalle los procedimientos y costos de los procesos de escaneado y reconocimiento óptico de caracteres (OCR por sus siglas en inglés) para obtener al final un texto con el formato adecuado para los programas Greenstone. También se describe la forma de crear y editar el material asociado a una colección.

Hemos procurado formular nuestras explicaciones de la manera más sencilla posible. Cualquier referencia a productos o empresas responde a un propósito meramente ilustrativo, y no presupone por nuestra parte valoración o preferencia alguna por tales productos en perjuicio de cualquier otro.

Conjunto de documentos

La serie completa de documentos comprende cinco volúmenes:

- La Guía de Instalación de la Biblioteca Digital Greenstone
- La Guía del Usuario de la Biblioteca Digital Greenstone
- La Guía del Programador de la Biblioteca Digital Greenstone
- La Biblioteca Digital Greenstone: del Papel a la Colección (*el presente documento*)
- La Biblioteca Digital Greenstone: uso del Organizador

Agradecimientos

Los capítulos dedicados al escaneado, el Organizador y demás información relativa a la creación de colecciones documentales colectivas sin fines de lucro son obra del Dr. Michel Loots, Gerente de la ONG Human Info y de HumanityCD, Dan Camarzan de Simple Words y el equipo que colabora con ambos desde Brasov (Rumania).

El programa Greenstone es fruto de la colaboración de muchas personas. Rodger McNab y Stefan Boddie son los principales arquitectos y programadores. También han contribuido David Bainbridge, George Buchanan, Hong Chen, Michael Dewsnip, Katherine Don, Elke Duncker, Carl Gutwin, Geoff Holmes, Dana McKay, John McPherson, Craig Nevill-Manning, Dynal Patel, Gordon Paynter, Bernhard Pfahringer, Todd Reed, Bill Rogers, John Thompson y Stuart Yeates. Otros miembros del proyecto Biblioteca Digital de Nueva Zelanda que proporcionaron asesoría y valiosas ideas para la concepción del sistema son: Mark Apperley, Sally Jo Cunningham, Matt Jones, Steve Jones, Te Taka Keegan, Michel Loots, Malika Mahoui, Gary Marsden, Dave Nichols y Lloyd Smith. También queremos dar las gracias a todos aquellos que contribuyeron a los paquetes de programas con licencias GNU incluidos en esta distribución: MG, GDBM, PDFTOHTML, PERL, WGET, WVWARE y XLHTML.

ÍNDICE

1 Introducción

Uno de los objetivos de la Biblioteca Digital Greenstone es hacer posible que instituciones como las universidades, organismos del sistema de las Naciones Unidas, organizaciones no gubernamentales, gobiernos y organizaciones no lucrativas puedan crear diversas colecciones de información y difundirlas por Internet o en forma de CD-ROM.

El proceso suele comprender los siguientes pasos:

i. Determinar los documentos que han de figurar en la colección.
ii. Obtener la autorización de los titulares de los derechos de autor para incluir esos documentos en la biblioteca digital.
iii. Escanear y reconocer por medio de OCR los documentos impresos que no estén disponibles en formato digital.
iv. Convertir todos los documentos en un formato que integre texto e imágenes y se pueda importar a Greenstone, de preferencia en formato HTML o Word de Microsoft, aunque también se cuenta con plugins que reconocen otros formatos con un nivel variable de precisión (véase la *Guía del Usuario de la Biblioteca Digital Greenstone*).
v. Etiquetar los capítulos, párrafos e imágenes de los documentos electrónicos.
vi. Dotar a la colección de la estructura idónea para que funcione como biblioteca digital.
vii. Crear la biblioteca digital utilizando los programas Greenstone.
viii. Editar y distribuir la colección en CD-ROM y/o distribuirla por Internet.

Para crear una colección digital es preciso que las publicaciones existan antes en formato digital. Los libros, revistas u otros documentos que sólo existan en forma impresa deberán ser escaneados, procesados y convertidos en un formato que la computadora reconozca (paso iii). El procedimiento más usual para ello es el reconocimiento óptico de caracteres (OCR por sus siglas en inglés), aunque a veces se opta por capturar íntegramente el texto. De este proceso tratan los capítulos 2 a 4 del presente manual.

El paso v) sirve para que, una vez finalizada la biblioteca, el usuario pueda seleccionar y ver en pantalla por separado las distintas partes de un documento. El paso vi) consiste en asignar atributos a cada documento, como por ejemplo categorías temáticas, palabras clave y datos bibliográficos con arreglo a los cuales se pueda ordenar y consultar la biblioteca. Estos pasos se explican en el Capítulo 5 del presente manual.

Asimismo, en el presente manual se tratan numerosas cuestiones relativas al procedimiento editorial y a la creación de una colección digital a partir de documentos impresos. Antes de seguir adelante, el lector debe plantearse las siguientes preguntas:

- ¿Cuál es el objetivo de mi colección?
- ¿A qué grupo de usuarios se dirige?
- ¿Cuán grande es dicho grupo? ¿Tiene dimensión local, regional o mundial?
- ¿Cuántos documentos tengo pensado incluir en la colección?
- ¿Cuántas páginas?

- ¿Cuánta información gráfica contienen?
- ¿Cabe subdividir la documentación en partes que sean de interés para un público reducido y otras que requieran una difusión más amplia?
- ¿Existen ya en forma electrónica los documentos?
- De ser así, ¿en qué formato? (Señalemos de paso que un archivo PDF no equivale automáticamente al texto completo en formato electrónico, pues a menudo contiene sólo la imagen de las páginas.)
- ¿A qué derechos de autor están sujetos los documentos?
- ¿Quién es el titular de los derechos de autor?
- ¿Hay otras organizaciones que se dirijan al mismo público?
- ¿Tengo deseos de colaborar con otros grupos?
- ¿De qué presupuesto dispongo para el conjunto del proyecto?
- ¿De qué recursos humanos (en personas-mes) dispongo para las labores de coordinación, edición, escaneado y programación?
- ¿De cuántas computadoras dispongo para el proyecto?
- ¿Cuántos CD-ROM tengo pensado distribuir?
- ¿Voy a distribuirlos gratuitamente o a venderlos?

2 Escáneres y escaneado

En el proceso de conversión de documentos impresos en una colección de biblioteca digital, el primer paso consiste en obtener imágenes digitales de todas las páginas de todas las publicaciones. La siguiente etapa es la del reconocimiento óptico de caracteres (OCR), proceso que requiere, para un resultado óptimo, imágenes de partida limpias y de buena calidad. Para el proceso de digitalización se necesita un escáner que pueda trabajar a una resolución de 300 ppp (puntos por pulgada). Gran parte del trabajo puede hacerse en blanco y negro, aunque deberá utilizarse un escáner capaz de obtener imágenes en color cuando el documento las contenga. La mayoría de las cubiertas de libro son en color, por lo que hay que escanearlas en forma de imagen fotográfica en color.

2.1 Escáneres

Existen escáneres de todos los precios, formas y tamaños, que pueden costar desde 100 dólares (un escáner plano) hasta más de 50.000 dólares (los grandes escáneres industriales de fabricantes como Bell & Howell[1]). Hay muchos sitios Web en los que se ofrece a la venta un amplio surtido de escáneres. Para encontrarlos basta con introducir la palabra "scanner" (escáner) en buscadores como Google, Altavista o Yahoo.

El formato de salida de una página escaneada es un archivo informático, por lo general en formato TIFF o Bitmap. El mejor formato es el TIFF IV comprimido. Una página normal, escaneada y convertida en este formato, ocupa sólo 50 Kb de memoria, mientras que una página equivalente en formato Bitmap no comprimido puede llegar a los 2 Mb.

Escáneres planos (o de sobremesa) de gama baja

Los escáneres de este tipo son los más económicos y difundidos. Existen muchas marcas: HP, Agfa, Acer, etc., con precios que van de los 100 a los 300 dólares. Con ellos pueden obtenerse imágenes tanto en blanco y negro como en color. El bajo precio de estas máquinas permite conectar cada computadora a su propio escáner.

Entre sus inconvenientes cabe citar la mediocre calidad del resultado, la lentitud con que trabajan, su escasa fiabilidad a temperaturas altas y la relativa frecuencia con que se averían. Es preciso escanear las páginas manualmente, una por una, colocándolas cuidadosamente en la placa de escaneado para que queden correctamente alineadas. De ahí que su productividad resulte baja. Aunque los fabricantes aseguran que se puede escanear una página en menos de un minuto, en la

[1] Todos los importes indicados en este documento se expresan en dólares estadounidenses y corresponden a las tarifas vigentes en 2001.

práctica rara vez se consiguen resultados superiores a las doce páginas por hora. Además, el proceso de escaneado monopoliza la computadora con la que se está realizando el trabajo.

Por todo lo dicho, estos escáneres sólo son útiles para realizar trabajos de escasa envergadura y pocas páginas: no más de 200 a 400 páginas al mes cuando se trate de un trabajo sistemático, y de 1.000 a 2.000 páginas para operaciones de carácter esporádico.

Escáneres de gama baja con alimentador de papel

Los escáneres de este tipo suelen costar entre 500 y 1.200 dólares. Ofrecen la posibilidad de escanear y procesar de 10 a 50 páginas de una vez, lo que evita que el operador tenga que estar continuamente pendiente de la máquina y aumenta la productividad hasta unas 150 a 200 páginas diarias. Estos escáneres son más robustos y gozan de una vida útil más larga antes de necesitar reparaciones (lo que suele ocurrir al cabo de 30.000 a 50.000 páginas).

Una de sus desventajas es que sólo pueden escanear una cara a la vez, lo que obliga a invertir el paquete de páginas y escanearlas de nuevo para obtener imágenes por ambas caras. Ello da lugar a frecuentes problemas, pues los alimentadores nunca funcionan a la perfección y a veces las páginas se atascan.

Estos escáneres son útiles para trabajos de 1.500 a 3.000 páginas mensuales.

Escáneres en color

Toda operación de escaneado conlleva siempre imágenes en color, lo que hace imprescindible un escáner capaz de procesarlas. Por regla general, menos del 5% de una publicación contiene imágenes en color, sin contar la cubierta. En consecuencia, será suficiente un escáner plano de gama baja como los mencionados más arriba. Es aconsejable elegir una máquina que pueda trabajar a una resolución de hasta 600 ppp.

Escáneres dúplex profesionales

Los escáneres profesionales son máquinas fiables y resistentes, capaces de tratar un gran número de páginas (normalmente entre 2.000 y 10.000 páginas diarias). Están provistos de una bandeja de alimentación automática con capacidad para 50 a 200 páginas. Los mejores y más rápidos son los dúplex, es decir, capaces de escanear simultáneamente las dos caras de una hoja.

Para utilizar un escáner dúplex profesional se requiere una computadora potente, dotada de un disco duro de 10 a 20 Gb de capacidad como mínimo. El precio de estos escáneres oscila entre los 5.000 y los 50.000 dólares. El escáner dúplex Canon DR-6020, por ejemplo, cuesta 5.000 dólares y puede trabajar con documentos impresos por ambas caras. Ofrece un rendimiento de unas 2.000 páginas diarias y un periodo de vida útil de 600.000 a 800.000 páginas. Los escáneres Bell & Howell y Fujitsu cuestan entre 10.000 y 50.000 dólares y gozan de un periodo de vida útil de muchos millones de páginas.

Los escáneres para microfichas cuestan entre 15.000 (por una unidad semimanual) y 80.000 dólares (por un escáner completamente automático).

Programas de escaneado

Cada escáner viene acompañado de su propio programa informático, que es necesario instalar en la computadora desde la que vaya a controlarse el escaneado. Algunos traen consigo una tarjeta controladora que se instala en la computadora para acelerar la digitalización.

2.2 Preparación de los documentos

Antes de escanear los documentos hay que prepararlos adecuadamente, eliminando posibles motas de polvo, secándolos si están húmedos, extrayendo clips y grapas y alisando las páginas dobladas o arrugadas.

Es preciso desmontar el lomo de los libros, cortándolo en línea recta y desprendiéndolo con precisión. A menudo habrán de encuadernarse de nuevo los libros procedentes de bibliotecas, en cuyo caso conviene extremar las precauciones al desmontar el lomo para facilitar la posterior encuadernación.

Cuando se trabaje con pocos documentos se puede retirar el lomo manualmente, con ayuda de una regla y una cuchilla. ¡Pero cuidado con los dedos! Para un mayor número de documentos merece la pena recurrir a guillotinas manuales especiales, y para grandes volúmenes, por ejemplo más de 20 documentos, recomendamos pedir permiso a una imprenta o copistería para utilizar su guillotina profesional. Recuérdese que es preciso extraer los clips y las grapas para no dañar las cuchillas.

2.3 El proceso de escaneado

Con el programa informático suministrado con el escáner se genera a partir de cada página una imagen electrónica, que se transforma en una imagen Bitmap o TIFF y se memoriza acto seguido en el disco duro, asignando a cada archivo un nombre normalizado. Una vez escaneados todos o una parte de los documentos de un lote empieza el proceso de OCR, del que puede ocuparse el operador del escáner o cualquier otra persona.

En general se necesita una resolución de 300 ppp, aunque a veces también resulte aceptable un valor de 200 ppp.

Control de calidad

La digitalización sirve para obtener una versión perfecta de las publicaciones en formato de texto o HTML mediante el proceso de OCR, o bien para crear archivos gráficos mejorados como los PDF. En ambos casos es de suma importancia que la imagen sea de buena calidad. En caso contrario los archivos gráficos resultan poco nítidos y consumen más memoria. La calidad influye sobremanera en el proceso de OCR: con imágenes de calidad mediocre la productividad puede caer hasta en un

40%. Sabiendo que el OCR suele representar más del 90% del costo total, se deduce que la calidad del escaneado es determinante para el costo final del proceso.

Es posible mejorar la calidad de un archivo TIFF adaptando el proceso de escaneado a cada tipo de papel mediante las opciones de ajuste que ofrece el programa del escáner. Un tipo de papel relativamente transparente requerirá parámetros más claros, el contraste deberá ajustarse en función de la calidad de la impresión, etc.

Ante todo conviene dividir el material en lotes de documentos que presenten similar calidad de papel e impresión, y después realizar pruebas de OCR con una muestra del primer lote para determinar los parámetros de ajuste idóneos. A continuación se escanearán todos los documentos de ese primer lote antes de continuar con el siguiente.

Convenciones para designar los archivos

Es preciso asignar a cada libro o documento un número de trabajo o código único, que a su vez dará nombre a la carpeta en la que se archiven todas las imágenes TIFF correspondientes a ese documento. Según el sistema operativo del que se trate (DOS, Windows, UNIX, LINUX, etc.), el nombre de un archivo puede comprender entre 8 y 128 caracteres, aunque es aconsejable limitarse a un máximo de 16 caracteres. Si se utilizan 8 caracteres, los cinco primeros servirían para identificar el documento, el siguiente sería una letra indicativa del código de idioma y los dos restantes caracteres indicarían el número de página. Por ejemplo: el identificador *u7548e12.tif* correspondería a la imagen TIFF de la página 12 de un libro escrito en inglés que tuviera por código *u7548e.*

Conviene asignar un directorio del disco duro a los trabajos de escaneado (por ejemplo con el nombre de *scanjobs*). Después se puede crear un subdirectorio para cada uno de los trabajos, en cuyo interior se creará un nuevo subdirectorio para cada publicación (en el ejemplo anterior, *u7548e*), donde se guardan todas las imágenes TIFF correspondientes a esa publicación, incluidas las imágenes en color.

2.4 Productividad y recursos necesarios

No hay que subestimar la carga de trabajo que suponen los procesos de escaneado y, sobre todo, de OCR. Es preferible considerar que ambos procesos son completamente independientes y elegir el procedimiento idóneo (desde el punto de vista económico y también práctico) para cada uno de ellos por separado.

Entre los aspectos que conviene tener en cuenta cabe señalar: la inversión necesaria en escáneres y computadoras, la existencia de los recursos humanos y el espacio necesarios, la formación del personal, los gastos salariales, el número inicial y total de páginas por escanear, los plazos en que ha de realizarse el trabajo y la posibilidad de subcontratarlo a terceros.

Costos del proceso de escaneado

Es importante decidir si se ha de invertir en un equipo de escaneado y asumir internamente esa labor o bien subcontratarla a una empresa especializada. Las principales consideraciones que deben tenerse en cuenta son:

- los plazos en que ha de realizarse el trabajo;
- el número total de páginas;
- los costos salariales correspondientes a las personas encargadas del escaneado.

Las personas a quienes se encomiende el escaneado deben estar muy motivadas y preparadas técnicamente, y tener muy clara la importancia de cumplir los criterios de calidad.

Una empresa especializada suele cobrar 0,06 dólares por página, a lo que hay que añadir los gastos de envío, que pueden ser de hasta 0,03 dólares por página cuando se remite el material de un país en desarrollo a un país desarrollado y de 0,015 dólares por página cuando se trata de un envío nacional.

En el Cuadro 1 se ofrecen estimaciones de los costos de escaneado por cuenta propia según el tipo de escáner que se utilice. Conviene tener en cuenta que estas cifras son aproximadas y corresponden más bien a órdenes generales de magnitud, basados en la experiencia de los autores. En las tres primeras columnas se presentan los costos laborales. En la primera se indica la productividad (en páginas mensuales) suponiendo que se trabaje a jornada completa. Dividiendo las horas mensuales de trabajo por la productividad en páginas/mes se obtienen los recursos necesarios en términos de horas de trabajo/página, guarismo que figura en la segunda columna (suponiendo que se trabajen 180 horas al mes).

Cuadro 1. Costos del proceso de escaneado

	Rendimiento (páginas/ mes)	Horas/página (180 horas/mes)	Costo/página (suponiendo 4 dólares/hora)	Adquisición del escáner (dólares)	Vida útil del escáner (páginas)	Páginas que se podrían subcontratar por el precio del escáner (a 0,06 dólares c/u)
Escáner plano	2.500	0,072	0,288	300	7.000	5.000
Escáner con alimentador de papel	8.000	0,0225	0,09	800	30.000	13.000
Profesional: dúplex de gama baja	40.000	0,0045	0,018	6.000	600.000	100.000
Profesional: dúplex de gama alta	150.000	0,0012	0,0048	50.000	8.000.000	833.000

Para determinar el precio por página se multiplican los costos salariales por hora totales (que dependerán de cada caso) por la segunda columna del Cuadro 1. En la tercera columna se indica, a título de ejemplo, el precio de escanear por cuenta propia suponiendo que se pague un sueldo de 4 dólares/hora, excluidos los gastos de inversión.

Estos cálculos presuponen que se procesa un número de páginas suficiente para justificar la adquisición de un escáner. En las tres últimas columnas del Cuadro 1 se ofrece información sobre los costos derivados del propio escáner. En la primera de ellas se indica el precio de adquisición de la máquina, en la siguiente el tiempo de vida útil que se le supone y en la última el número de páginas que podrían subcontratarse por el precio del escáner, contando una tarifa de 0,06 dólares/página.

Hay, desde luego, muchos otros factores que influyen en la decisión de adquirir o no un escáner: la disponibilidad de fondos suficientes, la necesidad de reducir al mínimo la dependencia para con terceras partes, el deseo de crear capacidades a escala local, la obligación que imponen las bibliotecas de escanear los libros en la propia localidad y no transportarlos, etc.

Las cifras del Cuadro 1 dan una idea aproximada del número de páginas necesario para justificar distintos niveles de inversión. No es frecuente que una institución u otra entidad necesite escanear 800.000 páginas. Con tales niveles de trabajo se plantean cuestiones más complejas que no vamos a tratar en estas líneas, como el mantenimiento del equipo o la posibilidad de recuperar gastos ofreciendo a terceros el servicio de escaneado.

Es tentador contemplar el desarrollo de la capacidad de escaneado como una actividad comercial, sobre todo en los países en desarrollo. Pero conviene tener muy en cuenta que no se trata de un proceso repetitivo. Una vez escaneado un documento, el cliente nunca cursará un nuevo pedido para repetir la operación, por muy buena que haya sido su relación de trabajo con la empresa. Desde un punto de vista comercial, se requiere un enorme trabajo de publicidad y comercialización. Desaconsejamos a cualquier ONG u otras organizaciones sin fines de lucro que se aventuren en este terreno sin haber procedido a ensayos exhaustivos y elaborado un minucioso plan comercial.

En conclusión, para escanear entre 10.000 y 50.000 páginas conviene plantearse la posibilidad de subcontratar el trabajo. Los cerca de 6.000 dólares que cuesta un escáner profesional de gama baja sólo se amortizan a partir de un volumen superior a las 100.000 páginas. Otra posible solución consiste en asociarse con otras instituciones (ONG o bibliotecas, por ejemplo) para adquirir colectivamente un escáner de ese tipo.

3 OCR: reconocimiento óptico de caracteres

Los sistemas de reconocimiento óptico de caracteres (OCR) transforman en texto una imagen escaneada. El punto de partida es una imagen digitalizada en formato TIFF o Bitmap, de la mayor nitidez y calidad posibles, y el resultado final un archivo de texto (generalmente en formato RTF o Word) o para la Web (formato HTML).

El proceso de conversión de un documento impreso en un archivo informático comprende las siguientes etapas:

1. escaneado;
2. análisis de la compaginación;
3. reconocimiento óptico de caracteres;
4. escaneado de ilustraciones y cuadros.

A lo largo del proceso se efectúan controles de calidad de los archivos resultantes y se memorizan éstos en el formato apropiado.

El mercado ofrece muchos y buenos programas de OCR, con precios que oscilan entre los 100 y los 400 dólares[2]. Entre muchos otros ejemplos cabe citar los siguientes:

- *Read-Iris* (http://www.readiris.com/)
- *Omnipage* (http://www.omnipage.com/)
- *Fine-Reader* (http://www.finereader.com/)

En los sitios Web de los fabricantes se ofrece toda la información necesaria, comprendida la lista de distribuidores locales. Los autores, de acuerdo con su experiencia, consideran que los programas de más fácil manejo son Fine-Reader y Omnipage. El primero, que cuesta unos 100 dólares, es el más barato y ofrece no sólo gran flexibilidad sino también el mayor repertorio de idiomas.

Es necesario decidir si se efectúan los procesos de escaneado y OCR internamente o se subcontratan a una empresa especializada. Hacerlo por cuenta propia exige disponer de un escáner, de un programa de OCR, de conocimientos técnicos en la materia y de personal muy motivado y atento a los requisitos de calidad.

---

[2] Recordemos que todos los importes están expresados en dólares estadounidenses de 2001 y corresponden a las tarifas vigentes en 2001.

3.1 El proceso de OCR

El proceso de OCR difiere según se utilice uno u otro programa de OCR, y cada uno de ellos exige un tiempo considerable de aprendizaje. En el manual de cada programa se exponen todos los detalles relativos al proceso. Hay cuatro aspectos que merecen especial atención: el control de calidad, los cuadros, las ilustraciones y los textos especiales como fórmulas, caracteres extranjeros, etc.

Control de calidad

Es preciso insistir en la importancia del control de calidad. Lo ideal es que esos controles estén a cargo de personas cuya lengua materna sea el idioma en que está escrito el documento o de gente con un excelente dominio del mismo. El perfil idóneo es el de alguien con estudios universitarios o secundarios. Conviene saber además que en este tipo de tarea los jóvenes suelen mantener un nivel de concentración superior.

Normalmente hay cuatro controles de calidad.

El primero se efectúa al mismo tiempo que el proceso de OCR. Cada programa tiene un verificador ortográfico incorporado que señala todas las posibles letras erróneas y muestra la imagen de la palabra entera para facilitar la comprobación y eventual corrección del error.

El segundo es un control general del texto una vez finalizado el proceso de OCR. Uno de los errores más frecuentes es la omisión de una página, un párrafo, los títulos de un capítulo, etc. Debe llevarse a cabo un repaso general para comprobar que no falta ninguna página. Es esencial asimismo comprobar los títulos, los encabezamientos de capítulo, los párrafos y los cuadros.

El tercer control es el ortográfico, para el que en general se utiliza Word de Microsoft porque su diccionario suele ser más completo que el de los programas de OCR. Importando el libro a un archivo Word y realizando un control ortográfico con este programa se puede detectar y corregir un mayor número de errores. Es indispensable añadir al verificador ortográfico cualquier palabra especialmente difícil o susceptible de generar una señal de error, así como los términos científicos y técnicos que abunden en el tipo de publicación con que se esté trabajando.

Finalmente, otra persona debe efectuar un último control del documento finalizado, tomando al azar fragmentos del libro completo y cerciorándose de que no haya errores o problemas con los cuadros, las ilustraciones, las leyendas o el aspecto general del documento. Sólo después de este último control puede considerarse que el libro está listo para su difusión electrónica.

Cuadros

Los cuadros suelen plantear dificultades a los programas de OCR. Controlar su contenido es además una labor ardua: contienen muchos dígitos, a veces con puntos y comas, y es fácil que las cifras acaben colocadas en la casilla equivocada. Es una tarea que exige concentración, dedicación,

un intenso trabajo de relectura, comprobaciones minuciosas y un buen control de calidad. Hay básicamente tres formas distintas de proceder.

La primera consiste simplemente en escanear los cuadros como si fueran imágenes en blanco y negro e insertarlos con este formato en el lugar correspondiente del documento. Esta es la solución más sencilla, pues no genera errores y no exige más tiempo que el necesario para crear la imagen. Pero consume más memoria que las dos restantes, y además la resolución obtenida no siempre basta para trabajar en la computadora con cuadros de gran tamaño: si se reduce todo el cuadro a los límites de la pantalla, la resolución es demasiado pequeña; si por el contrario el cuadro desborda la pantalla, el usuario debe desplazarse para ver todas las columnas y filas, con lo que pierde visión de conjunto.

El segundo método es la copia manual: crear un nuevo cuadro con el mismo número de filas y columnas y copiar los valores correspondientes a cada casilla, carácter por carácter.

La tercera solución consiste en someter el cuadro al proceso de OCR. Aunque este procedimiento ahorra tiempo en comparación con el manual, la probabilidad de error es más alta. A veces las columnas quedan fusionadas, o el programa es incapaz de reconocer los puntos y comas.

Ilustraciones

Las ilustraciones contenidas en una publicación corresponden en general a tres grandes tipos de imagen:

- ilustraciones en blanco y negro, sin tonos intermedios;
- fotografías en blanco y negro;
- fotografías en color.

Las ilustraciones en blanco y negro deben escanearse en modo “dibujos de líneas simples” y guardarse en formato GIF o PNG. Para las fotografías en blanco y negro conviene utilizar el modo “escala de grises” y guardar el resultado en archivos GIF o JPEG. En cuanto a las fotografías en color, es preciso escanearlas en modo “color” y guardarlas en archivos JPEG. En términos generales, el formato JPEG de calidad media ofrece una resolución suficiente.

Las ilustraciones suelen consumir gran parte del espacio que ocupa una colección en el disco duro o el CD-ROM. De ahí la importancia de lograr para cada imagen la mayor claridad y visibilidad junto con el menor tamaño posible. Para ahorrar espacio cabe la posibilidad de prescindir de algunas imágenes o de todas ellas cuando no sean necesarias para entender el texto.

Las ilustraciones deben escanearse por separado, una por una. Para denominar los archivos gráficos recomendamos un nombre compuesto por los cinco o seis primeros caracteres utilizados para designar el documento seguidos del número de la página en que se encuentre la ilustración. Una alternativa, suponiendo que haya un directorio para cada documento, consiste simplemente en utilizar la letra *p* [*picture*] seguida del número de la página. Cuando en una misma página haya varias ilustraciones, bastará con añadir una letra *a, b, c* ... al nombre del archivo. Por ejemplo, a

una imagen JPEG que aparezca en la página 36 de la antedicha publicación *u7548e* corresponderá un archivo llamado *u7548e36.jpg* o *p36.jpg*.

Una vez escaneadas las imágenes, se pueden aplicar programas de procesamiento por lotes para modificar las dimensiones o mejorar la definición de todas las imágenes a la vez.

Textos con características especiales

Muchos documentos contienen elementos que conviene tratar aparte (caracteres especiales, fórmulas, páginas especialmente dificultosas, etc.). Los caracteres especiales suelen provenir de idiomas distintos u ostentar marcas diacríticas. En tal caso hay que seleccionar el idioma del que se trate en la opción "idioma" del programa OCR. Las fórmulas deberán reproducirse manualmente, lo que a veces es imposible con un programa de OCR, en cuyo caso hay que recurrir a un procesador de texto como Word de Microsoft. Las páginas de las que no pueda obtenerse una imagen nítida, ya sea por la complejidad del texto o por el mal estado en que se encuentren, deberán ser reproducidas manualmente.

3.2 Productividad y recursos necesarios

Como hemos dicho, no hay que subestimar la dificultad del proceso de OCR. Aunque conviene estudiar separadamente las alternativas económicas y prácticas del proceso de escaneado y del de OCR, ambos plantean interrogantes parecidos: la inversión necesaria en computadoras, la disponibilidad de personal y de capacidad de gestión, la formación del personal, los costos salariales, el número total de páginas que deben tratarse y las posibilidades de subcontratar el trabajo a terceros.

Esta sección se basa en la experiencia de los autores en el trabajo de OCR en Bélgica, Rumania y la India. Todos los ejemplos, cálculos y cifras que aquí se exponen corresponden a una situación ordinaria: documentos de dificultad normal (con cuadros e ilustraciones) como los que pueden encontrarse en la mayoría de los archivos o bibliotecas, resultados de muy buena calidad y trabajo a medio o largo plazo.

Trabajo intensivo de OCR

El OCR es un proceso difícil, que exige gran concentración y destreza. Antes de alcanzar un nivel óptimo de rendimiento y calidad, el operador necesita un periodo de aprendizaje de unas seis semanas.

Los mejores resultados y la productividad más alta se consiguen por lo general durante las primeras horas de trabajo. Al cabo de tres horas la productividad baja con rapidez, quizá hasta un 50% del nivel inicial. Al cabo de seis horas, la mayoría de la gente se encuentra muy cansada.

Algo parecido ocurre durante las primeras semanas de trabajo, en las que todo el mundo alcanza una productividad bastante elevada. Posteriormente, sin embargo, hasta dos tercios de los operadores de OCR empiezan a sentirse aburridos y descontentos. A la larga esas personas acaban

abandonando el trabajo o rindiendo poco en términos de calidad y productividad. Incluso los que superan el periodo crítico de tres a cinco semanas y se integran en el equipo de trabajo suelen renunciar y partir en busca de una mejor ocupación al cabo de 6 a 12 meses.

Las observaciones sobre el personal que formulamos en la sección 3.1 son especialmente aplicables al trabajo intensivo de OCR. Los controles de calidad resultan mejores cuando corren a cargo de hablantes nativos o profundos conocedores del idioma en cuestión. En general los jóvenes pueden mantener un nivel de concentración superior al de las personas mayores en las labores de OCR. La experiencia demuestra que las personas de entre 18 y 23 años de edad tienden a adaptarse mejor a ese cometido que las mayores de 25 años.

Por último, considerando lo aburrido que puede resultar el trabajo de OCR, la motivación y un constante prurito de calidad son elementos de excepcional importancia.

De todo lo dicho se desprenden las siguientes directrices generales sobre el proceso de OCR:

- Los jóvenes de entre 18 y 25 años de edad son los más aptos para este tipo de trabajo.
- Dado que las primeras horas son siempre las más productivas, conviene organizar turnos de trabajo a tiempo parcial o, en su defecto, encomendar la labor a jornada completa a las personas más motivadas y con mayor capacidad de concentración.
- Después de tres a cinco semanas de actividad, dos tercios de los operadores tienden a renunciar o a sentirse hastiados. Ello se traduce en un descenso de la calidad y la productividad en las últimas semanas.
- Es preciso un suministro periódico de trabajo para justificar la necesaria formación del personal, mantener la concentración y conservar alta la moral del equipo.

Objetivos asequibles de productividad

En el Cuadro 2 se presentan las cifras más frecuentes de productividad en el trabajo de OCR. Teniendo en cuenta que puede tratarse de documentos de todos los tamaños y niveles de calidad, estas cifras parten del supuesto de que el conjunto de documentos contiene un número promedio de ilustraciones y cuadros (por ejemplo una ilustración y un cuadro de 5x5 cada ocho páginas), que las ilustraciones son de calidad entre media y alta (recordemos que ello depende de la calidad del escaneado) y que los operadores de OCR dominan el idioma en que está escrito el documento.

Cuadro 2. Productividad en el proceso de OCR

	Horas de trabajo/día	Páginas/día	Páginas/mes
Formación inicial (seis semanas)	3	6	120
Nivel óptimo de productividad	3	9	150 a 200
	7	28	500 a 600

En el Cuadro 2 se distingue entre las estadísticas de personas en periodo de formación y las de quienes han alcanzado su nivel óptimo de productividad. Si un miembro del personal administrativo dedicara tres horas diarias al trabajo de OCR, su rendimiento sería de entre 180 y 200 páginas al mes. Un operador a jornada completa bien formado, con gran capacidad de concentración y escrupulosa atención a los criterios de calidad, en cambio, podría alcanzar una productividad de entre 500 y 600 páginas al mes.

Sin embargo, con páginas de especial dificultad y escasa calidad, con abundantes cuadros o columnas, se obtienen cifras muy inferiores (quizá de 300 a 400 páginas mensuales a jornada completa).

Supongamos que el costo salarial de un operador a jornada completa muy aplicado y motivado asciende a 400 dólares mensuales, y que los gastos generales (gastos de gestión, computadoras, espacio de oficina, instalaciones, etc.) suponen otros 300 a 400 dólares mensuales por persona. En tal caso, el costo del proceso de OCR viene a ser de 1,2 a 1,6 dólares por página. Si además se toma en cuenta el periodo de formación, el volumen total, el lapso de tiempo considerado y los eventuales costes de la suspensión de las operaciones cuando falte el trabajo, el costo asciende a un valor entre 1,5 y 2,5 dólares por página.

Conviene comparar los costos del proceso de OCR efectuado por cuenta propia con los de la subcontratación a una empresa especializada. Estas empresas suelen cobrar entre 1,5 y 4 dólares por página, incluyendo las ilustraciones y los cuadros. Human Info/Simple Word, que posee una unidad de este tipo en Rumania, aplica tarifas especiales para las organizaciones humanitarias sin fines de lucro (entre 1,2 y 2 dólares por página). Puede solicitarse información o asesoramiento escribiéndonos a la dirección: scanning@humaninfo.org.

3.3 Alternativas al proceso de OCR

En las siguientes líneas exponemos dos posibles alternativas al OCR.

Mecanografiado manual

La primera posibilidad, que además elimina buena parte del escaneado, consiste en mecanografiar de nuevo los documentos con un programa de tratamiento de texto. Utilizando este procedimiento hay que escanear únicamente las ilustraciones y la cubierta (y no las restantes páginas), lo que hace innecesario disponer de un escáner y un programa de OCR potentes.

No es preciso que los operadores entiendan el texto. Sólo tienen que ser buenos mecanógrafos y reproducir exactamente lo que ven. Dado que este proceso suele generar errores, a menudo se utiliza el doble mecanografiado para detectarlos y corregirlos. Este método requiere que dos personas mecanografíen independientemente el mismo documento, después de lo cual un operador provisto del texto original compara ambas versiones electrónicas palabra por palabra, con ayuda de un programa informático especial. Se parte de la premisa de que si una misma palabra ha sido escrita dos veces por separado de la misma manera, será correcta. Pero ello no siempre es así, y cuando se quiere trabajar con la máxima fiabilidad se recurre al triple mecanografiado.

Teniendo en cuenta que el uso de un programa de OCR entraña el de computadoras de gran potencia, la ventaja básica de este método es que prescinde del OCR y por lo tanto permite utilizar computadoras más antiguas, sencillas o de segunda mano, lo que supone un ahorro considerable. Además, esta labor requiere trabajadores menos especializados. En cuanto a sus inconvenientes, éstos residen en el periodo de formación (de al menos dos meses) que se necesita y en la abundancia de errores que suelen darse con un proceso de mecanografiado único, lo que obliga a trabajar por duplicado o triplicado.

Los costos de este procedimiento dependen exclusivamente del nivel salarial. Los mecanógrafos de países en desarrollo suelen cobrar unos 150 dólares mensuales. Su productividad oscila entre 20 y 30 páginas diarias, lo que equivale a 400 páginas mensuales, comprendidas las ilustraciones. Suponiendo que se trabaje por duplicado, los costos salariales suman en total 300 dólares al mes, sin contar los gastos generales.

Archivos gráficos

Una alternativa sumamente barata al proceso de OCR consiste en utilizar simplemente una versión gráfica en PDF de las páginas del documento, lo que reduce los costos a unos 0,1 dólares por página (una pequeña fracción de lo que costaría un proceso de OCR).

Una vez concluido el escaneado y creados los archivos TIFF, se utiliza un convertidor automático (en general Acrobat o Photoshop de Adobe) para convertir en formato PDF todos los archivos TIFF correspondientes a las páginas del libro.

El problema es que en esos archivos no se pueden efectuar búsquedas y que además son bastante pesados (por lo general 50 Kb por página, con un margen de variación del 20% según la calidad del archivo TIFF original).

La descarga de un archivo gráfico PDF es un proceso lento, a veces imposible o de precio prohibitivo en los países en desarrollo. Esos archivos caben rara vez en un disquete y no admiten operaciones de manipulación del texto como la de “cortar y pegar”.

Sólo se optará por esta solución cuando se carezca del presupuesto necesario para un proceso de OCR o cuando se trate de documentos destinados a un público poco numeroso y provisto de una conexión a Internet de bajo costo y alta velocidad.

3.4 Combinación de escaneado y OCR

La mayoría de los programas de OCR pueden escanear una página y efectuar inmediatamente el reconocimiento óptico, a condición de que el escáner esté conectado directamente a la computadora que ejecuta el programa. Aunque escanear y efectuar el OCR página a página es un método razonable cuando se trabaja con pocos documentos, resulta muy largo para trabajos más voluminosos y continuos.

Esta solución es adecuada para cantidades entre 100 a 150 página al mes. Para tratar volúmenes superiores, en cambio, es más rápido y eficaz escanear en primer lugar el documento y aplicar después el proceso de OCR a todas las páginas de una sola vez.

4 De 1.000 a 100.000 páginas en tres ejemplos

4.1 Una colección de pequeñas dimensiones: de 500 a 1.000 páginas

La mayoría de las ONG tienen un volumen de 500 a 1.000 páginas por escanear. Si disponen de voluntarios motivados pueden asumir por cuenta propia el proceso de OCR.

ESCANEADO

El primer paso consiste en escanear las publicaciones para generar un archivo TIFF de alta calidad para cada página y una imagen bitmap independiente (ya sea de dibujo de líneas simples, escala de grises o color) para cada ilustración. Suponiendo que deban escanearse 1.000 páginas, ello puede equivaler a cerca de un mes de trabajo a tiempo parcial (sólo para el escaneado). Los archivos TIFF ocuparían entre 60 y 80 Mb de espacio en el disco duro, por lo que es aconsejable utilizar un CD-ROM para dar cabida a esos archivos. Un escáner plano de precio reducido (entre 100 y 300 dólares) basta para realizar ese trabajo, del que puede ocuparse un voluntario después de la jornada laboral o durante los fines de semana, ya sea en la oficina o en casa.

OCR

La segunda etapa es la del proceso de OCR, que se encomendará a otro voluntario, o equipo de voluntarios, con buenos conocimientos lingüísticos y de corrección ortográfica. Cabe la posibilidad de repartir los archivos TIFF entre varias computadoras o bien de utilizar una sola máquina para la totalidad del trabajo. Por lo general se requieren entre cinco y seis meses de trabajo a tiempo parcial (a razón de 20 horas semanales, por ejemplo) para convertir 1.000 páginas en documentos impecables en formato Word o HTML.

SUBCONTRATACIÓN

Una posibilidad alternativa es la de subcontratar los procesos de escaneado y OCR. La conversión de todos los documentos en archivos Word y HTML impecables costaría probablemente entre 1.500 y 2.000 dólares.

4.2 Todas las publicaciones de una organización: 5.000 páginas

Los archivos de muchas organizaciones de mayor tamaño pueden contener unas 5.000 páginas de libros (en catálogo o agotados), revistas, boletines, documentos, etc.

ESCANEADO

Se trata de un volumen excesivo para un escáner plano, lo que deja dos opciones: subcontratar el trabajo (contando unos 400 dólares por 5.000 páginas) o adquirir un escáner con alimentador de papel (aproximadamente 900 dólares). Otra alternativa es que varias instituciones u ONG adquieran conjuntamente un escáner más caro (6.000 dólares, divididos por el número de participantes). Las 5.000 páginas en formato TIFF ocuparían entre 300 y 400 Mb en el disco duro. Señalemos de nuevo la conveniencia de utilizar un CD-ROM para guardar esos archivos.

OCR

A continuación hay que ocuparse del proceso de OCR, que puede encargarse a otro voluntario, o equipo de voluntarios, diestro en técnicas de OCR y corrección ortográfica. Como en el caso anterior, es posible utilizar varias computadoras o una sola para esta tarea. La conversión de 5.000 páginas en archivos Word o HTML impecables exigiría entre 25 y 30 meses de trabajo a tiempo parcial (a razón de 20 horas semanales), lo que en la práctica descarta el empleo exclusivo de voluntarios porque lleva demasiado tiempo y requiere un uso excesivo de la computadora. Para concluir el trabajo en un plazo razonable y con un buen nivel de calidad habría que pagar a los voluntarios, supervisar su rendimiento y la calidad de su labor, proporcionarles espacio adecuado, etc.

Una posibilidad alternativa es la de crear archivos gráficos PDF, que ocuparían entre 300 y 400 Mb de memoria y resultarían más difíciles de descargar de Internet.

SUBCONTRATACIÓN

Otra alternativa es la de subcontratar los procesos de escaneado y OCR, lo que costaría probablemente entre 7.500 y 10.000 dólares.

4.3 Una pequeña biblioteca: 100.000 páginas

Otras entidades de mayor envergadura, universidades, gobiernos o bibliotecas especializadas podrían tener una biblioteca entera por digitalizar, algo así como unas 100.000 páginas. Lo primero que se debe tener en cuenta es la situación de las publicaciones en materia de derecho de autor: bien están incluidas en el dominio público o bien hay que obtener permiso explícito de los titulares de los derechos para poder digitalizarlas. Conviene cerciorarse asimismo de que los documentos no existen ya en formato electrónico.

ESCANEADO

100.000 páginas son demasiadas para un escáner con alimentador de papel, por lo que caben dos opciones: subcontratar el trabajo (a unos 8.000 dólares las 100.000 páginas) o adquirir, conjuntamente con otras instituciones u ONG, un equipo de mayor calidad y precio (6.000 dólares, divididos entre los participantes). 100.000 páginas en formato TIFF ocuparán entre 6 y 8 Gb en el disco duro. Lo ideal es crear copias de esos archivos en discos CD-ROM.

OCR

La segunda etapa es el proceso de OCR (en el caso de documentos menos utilizados, también cabe la posibilidad de crear archivos PDF). Convertir 100.000 páginas en archivos Word o HTML impecables llevaría entre 500 y 700 meses de trabajo a tiempo parcial, lo que a todas luces descarta el uso de voluntarios para esta tarea, más propia de profesionales.

Para ahorrar costos se pueden convertir en PDF algunas de las páginas menos utilizadas (digamos el 80%, u 80.000 páginas), reservando los formatos Word o HTML para las 20.000 páginas restantes. Los archivos PDF ocuparían entre 4 y 6 Gb de espacio y no sería fácil descargarlos de Internet, pero en cambio saldrían baratos si los creara una empresa especializada (sólo 0,2 dólares por página, lo que supone un costo total de 16.000 dólares). Utilizando voluntarios para crear 80.000 archivos PDF a partir de archivos TIFF mediante programas de conversión como Acrobat de Adobe se precisarían 10 a 20 meses de trabajo a tiempo parcial con una computadora de gran potencia.

SUBCONTRATACIÓN

Una posible alternativa es la de subcontratar el trabajo. Suponiendo que se mantuviera la mencionada proporción del 80% en PDF y el 20% en HTML, los archivos PDF costarían unos 16.000 dólares y los archivos HTML entre 30.000 y 40.000 dólares, con lo que el presupuesto total ascendería a unos 50.000 dólares. Si se sometieran todos los documentos a un proceso de OCR, convertir toda la colección en archivos Word y HTML impecables costaría entre 150.000 y 200.000 dólares.

5 Creación de una colección digital

Hay tres aspectos importantes que conviene tener en cuenta a la hora de crear una colección digital. En primer lugar es preciso organizarla. A mayor volumen de contenido, mayor necesidad hay de índices y sistemas potentes de búsqueda, indispensables cuando la colección supera las 3.000 a 5.000 páginas. En segundo lugar, deben prevalecer las necesidades del usuario final. Es preciso identificar los grupos que usarán la colección y establecer un proceso de consulta periódica con ellos. En tercer lugar, el presupuesto disponible determinará qué tanto se puede hacer.

5.1 Métodos para crear colecciones

Abundan los ejemplos de excelentes CD-ROM elaborados siguiendo el modelo de una página Web, en la que por medio de hipervínculos se insertan y enlazan entre sí documentos en formato HTML, PDF o Word. El uso de hipervínculos, marcos y grupos estructurados, palabras clave, índices y demás elementos de este tipo hace fácil y atractiva la navegación. Estos sistemas funcionan bien con volúmenes de unos cuantos miles de páginas, pero a partir de 3.000 a 5.000 páginas es importante que la colección esté bien organizada y ofrezca un dispositivo de búsqueda potente. Ahí es donde el programa Greenstone puede resultar de utilidad.

El programa Greenstone crea una biblioteca digital estructurada y provista de un poderoso buscador y un mecanismo de recuperación. Es posible indexar hasta 150,000 páginas en un solo CD-ROM, que además puede funcionar como servidor de Internet. Greenstone es un software de código fuente abierto y por lo tanto puede obtenerse gratuitamente bajo las condiciones estipuladas en la Licencia Pública General de GNU.

En los manuales de consulta que acompañan al programa se explica cómo crear colecciones Greenstone. Básicamente existen tres formas de hacerlo.

- Con la interfaz de bibliotecario
- Con el programa de recopilación, también conocido como Colector
- Crearlas desde la línea de comandos.

El primer método corresponde a la interfaz de bibliotecario descrita la *Guía del Usuario de la Biblioteca Digital Greenstone* (Capítulo 3, "Creando colecciones Greenstone"). Ésta es una herramienta interactiva para la creación de colecciones que permite reunir grupos de documentos, importar o asignar metadatos e integrarlos a una colección. El segundo método es el subsistema de recopilación descrito en el Capítulo 4 de la *Guía del Usuario* llamado "Colector". Se trata de una herramienta que aparece en versiones anteriores y ofrece una alternativa para la creación de colecciones de páginas web u otros documentos, guiándolo a través de una secuencia de páginas Web interactivas que solicitan la información conforme vaya siendo necesaria. No obstante, no proporciona ninguna forma de agregar metadatos a los documentos y, debido a su interfaz Web, no es realmente adecuada para colecciones cuya construcción requiera más de unos cuantos minutos. El tercer método es ejecutar los programas que permiten construir la colección directamente desde la línea de comandos que se encuentra en el Capítulo 1 de la *Guía del Programador de la*

Biblioteca Digital Greenstone. Este método le ofrece una mayor flexibilidad para ejecutar los programas de manera individual y le ahorra los pasos intermedios que quizá fueran deseables para colecciones que requirieran de muchas horas en su construcción. También necesitará leer el Capítulo 2 con el fin de aprovechar todo el poder que le ofrece Greenstone para la creación de colecciones avanzadas.

Existe una cuarta herramienta para crear y editar el material asociado a una colección llamada Organizador. Sin embargo, sus funciones han sido sustituidas por las de la interfaz de bibliotecario mencionada arriba. Este método se describe en el documento titulado *Uso del Organizador*.

5.2 Aprendiendo a usar la interfaz en siete pasos y 15 minutos

La mejor forma de conocer las características y el funcionamiento de la interfaz de bibliotecario es crear una pequeña biblioteca de prueba. Si dispone de 15 minutos, por favor siga los pasos que se indican a continuación y así usted obtendrá una mejor comprensión de este programa.

Antes de empezar lo primero que deberá hacer es instalar Greenstone (vea la *Guía de Instalación*), el cual incluye una colección de muestra en formato DLS y sus archivos fuente. **Recuerde, si desea añadir a su colección cualquiera de los 140 documentos de la colección DLS completa (en vez de sólo los 11 de esta colección de muestra) también deberá instalar la DLS como una de las bibliotecas Greenstone de muestra.** Las colecciones de muestra y DLS se instalarán en *C:\Program Files\gsdl\collect*, en los subdirectorios *demo* y *dls* respectivamente. Si anteriormente usted ya instaló Greenstone sin la colección DLS y desea instalarla ahora, sólo tiene que insertar nuevamente el CD-ROM Greenstone y añadir la colección. No es necesario desinstalar Greenstone primero.

Le sugerimos que imprima las instrucciones que aparecen a continuación y las siga paso a paso :

1. Para iniciar la interfaz bajo Windows seleccione *Biblioteca Digital Greenstone* en la sección de *Programas* del menú de *Inicio* y elija *Interfaz de Bibliotecario*. Si en vez de Windows usted está usando UNIX escriba:

   ```
   cd ~/gsdl
   cd gli
   ./gli.sh
   ```

   donde *~/gsdl* es el directorio que contiene su sistema Greenstone.

2. Seleccione *Nueva* en el menú Archivo que se encuentra en la barra horizontal en la parte superior de la ventana. Dele un título, por ejemplo “Mi primera colección” y escriba su dirección de correo electrónico y una breve descripción de la colección. En el menú “Basar esta colección en” elija “colección de muestra Greenstone” o “Subconjunto de la Biblioteca para el Desarrollo” (DLS por sus siglas en inglés). El efecto es el mismo, ya que ambas colecciones tienen la misma estructura.

3. Añada algunos documentos de la colección de muestra (o de la colección DLS si está instalada) a su nueva colección. Para ello haga doble click en la carpeta de *Colecciones Greenstone* en el cuadro izquierdo y a continuación haga doble click en la colección que prefiera. Los documentos que hay en ella aparecerán en pantalla. Seleccione uno, arrástrelo y colóquelo en el cuadro derecho. Este panel representa la colección que está construyendo. Elija varios documentos y arrástrelos uno por uno o seleccione y arrastre varios de ellos al mismo tiempo de la manera normal.

4. Agregue algunos de sus propios documentos que no estén en la colección de muestra o en la DLS. Cierre la carpeta de *Colecciones Greenstone* en el cuadro izquierdo y haga doble click en la carpeta *Local Filespace (Espacio de archivo local)*. Vaya a un directorio que contenga algunos documentos (p. ej. pequeños archivos de Word o HTML) y arrastre unos cuantos de ellos al cuadro derecho para incluirlos en su colección.

5. Añada metadatos a los documentos de su colección. Hasta este momento usted ha estado operando en el panel indicado por medio de la pestaña *Gather (Reunir)* que se encuentra debajo de la barra de menús horizontal en la parte superior de la ventana. Haga click en la pestaña *Enrich (Enriquecer)* que se encuentra a un lado. Los documentos de su colección aparecerán ahora en el cuadro del lado izquierdo. Haga click en uno y examine los metadatos asociados a él que se muestran en la tabla *"Element ... Value" (Elemento ... Valor)* en la parte superior derecha. Use el cuadro que está debajo para cambiar los valores individuales seleccionando el elemento que desee y escogiendo un valor existente de la lista o escribiendo un nuevo valor en el recuadro que se encuentra cerca de la parte inferior. Añada los metadatos *Título*, *Organización* y *Palabra clave* para cada uno de sus documentos que quiera poner en la colección. Después de escribir cada valor usted necesitará hacer click en *"Append" (Agregar)* para guardar dicho valor.

6. Haga click en la pestaña *Create (Crear)* para salir del modo *Enriquecer* y crear su nueva colección. Haga click en el botón *Build Collection (Construir la colección)* que se encuentra en la parte inferior. Conforme la computadora va construyendo la colección usted recibirá información sobre lo que está haciendo.

7. Una vez que haya terminado haga click en la pestaña *Preview (Vista previa)* para ver la colección desde el interior de la interfaz de bibliotecario. Revise las listas de *títulos de la "a" la "z"*, *organizaciones* y *cómo hacer* para asegurarse de que sus documentos han sido incluidos en la colección. Asimismo cuando visite su página principal de Greenstone usted encontrará que la colección ha sido instalada como una de las colecciones regulares.