Grupo de Trabajo BUC-Google

Documentos BUC

Serie: Informes

Resumen en español del documento *Technical Specifications for Library Partners (12/01/2006). Revision 4.0*

4 de enero de 2007

Preparado por:

Eugenio Tardón (Subdirección de Sistemas de Información Bibliográfica)

Manuela Palafox (Servicio de Edición Digital y Web)

1. INTRODUCCIÓN

Este documento es confidencial y no puede distribuirse a ninguna persona que no pertenezca al Grupo de Trabajo Google de la Biblioteca Complutense. El documento no pretende traducir el documento original, sino servir como documento de trabajo resumen. Contiene las especificaciones sobre captura, procesamiento y distribución de los libros escaneados por Google.

2.CONFIDENCIALIDAD

Este documento es confidencial, según se manifiesta en el apartado 1.

3. CAPTURA DE IMÁGENES

3.1. Material escaneable

Tamaño de los libros susceptibles de digitalizar: Máximo de 45x30 cm. y 8,8 cm. de lomo. Se escanea de cubierta a cubierta. Hay que cortar las páginas de los libros intonsos.

Google no escaneará libros frágiles, ni materiales sueltos, libros con lomos dañados, mapas, manuscritos, fotos, periódicos o libros sin códigos de barras.

3.2. Formato y resolución de escaneado

Formato de entrega: jpeg2000, tras transcodificar de jpeg a 8 bits por canal. A 200 ppi (si es > 29x19) y a 300 ppi (si es <29x19). Google escanea una o dos veces los libros.

4. INVENTARIO Y GESTIÓN DEL PROCESO

Los libros escaneados deben tener todos un código de barras único que actúa como identificador para:

1) Seguir las operaciones de escaneado

2) Identificar los ficheros de imágenes entregados a la Biblioteca

3) Enlazar un volumen con la información bibliográfica

4) Enlazar un volumen con notas de producción asociadas al escanear

La Biblioteca debe comunicar a Google su formato de código de barras y enviar ejemplos, según indica el apéndice A.

5. INFORMACIÓN BIBLIOGRÁFICA Y ENTREGA DE METADATOS

Hay que proporcionar a Google información bibliográfica y del ejemplar para cada volumen: en formato MARC 21 XML y UTF-8. Cuanto más completo sea el registro bibliográfico, mejor. Hay que incluir la información del ejemplar en la etiqueta 955 (955.b para el código de barras, 955.v número de volumen, utilizando otros subcampos para añadir la información que queramos).

Se puede enviar todo el catálogo o partes, con al menos de 24 horas antes de la recogida de los libros para escanear. El tamaño máximo de cada fichero será de 64 MB (pueden enviarse varios ficheros en un único *zip)*.

Los nombres de los ficheros siguen la convención: Complutense (Socio)-AAAAMMDD-HHMM.xml. Ejemplo: *<partnername>-200610102-0300.xml*

Distribución de los ficheros y registros:

- Todos los ficheros siempre deberán enviarse al mismo servidor y directorio. Google automáticamente extraerá información de los servidores de la biblioteca. Google soporta los protocolos HTTP y HTTPS para este tipo de acción.

- Google enviará informes con listas de los códigos de barras de los libros escaneados de los que no disponen del registro bibliográfico y del ejemplar.

6. PROCESAMIENTO DE IMÁGENES

Google busca lograr imágenes de páginas de libros limpias, normalizadas y bien alineadas. Para ello efectúa las tareas de:

- ***dewarping*** (eliminar curvaturas de páginas)
- ***cropping*** (recortar la página)
- ***de-skewing*** (alineamiento horizontal del texto)
- ***normalization*** (normalización de tonos de fondo y primer plano del texto)
- ***color processing*** (procesamiento independiente de las páginas en color, evitando la normalización de tonos)

7. OPCIONES DE ACCESO A LOS LIBROS

Google ofrece dos mecanismos o interfaces de acceso:

a) **Interfaces públicas**

- Universal, via Google Book Search. GBS es una interfaz de acceso universal (books.google.com) que permite acceder a los libros escaneados. Cada obra tiene un *PURL* según el ejemplo: http://books.google.com/books?vid=[PartnerName][Barcode]
- Restringido y personalizado (logo y color UCM) para usuarios complutenses. Es el interfaz de GBS para la UCM denominado *hosted solution*. Contiene PURL según el ejemplo: http://books.google.com/books/[PartnerName]?vid=[PartnerName][Barcode]

b) **Interfaz administrativo** (Apéndice B). El GRIN (Google Return Interface) es una interfaz Web que se utiliza para descargar los libros escaneados desde Google a la Biblioteca. Es un mecanismo para compartir información sobre los contenidos. Esta interfaz permite a las bibliotecas ver el estado de todos los volúmenes que han entrado en la operación del escaneado, descargarse ficheros convertidos como archivos encriptados, etc. Se pueden hacer *scripts* utilizando utilidades HTTP, tales como *wget*.

8. ESPECIFICACIONES DE LOS FICHEROS DE IMÁGENES

Formatos de distribución de los ficheros de imágenes

Google ofrece varios formatos. Cada socio solo puede elegir uno.

1. JPEG, a 300 ppi. Pesos medios: 200-500 kb pág sólo texto y 300-700 con imágenes. Color 24 bits.
2. JPEG2000, a 300 ppi. Pesos menores entre 50-75% al anterior. Color 24 bits, 8 capas. 50980 como *slope rate distortion*.
3. TIFF G4 (para página de sólo texto) y JPEG o JPEG2000 (para no texto). El texto en TIFF G4 interpolado a 600 ppi. Tamaño normal de 50 KB por página.
4. Profundidad de color general: texto (1 bit), escala de grises (8 bits), color 24 bits.

Convenciones para los nombres de ficheros

Al escanear un libro, Google genera un directorio para cada ejemplar cuyo nombre es el identificador del código de barras. En ese directorio sitúa todas las páginas del libro. Ejemplo:

barcode/00000001.jpg	front cover
----------	----------
00000090.tif	back cover

Es decir, el código de barras, un guión y un dígito de ocho caracteres para las páginas.

Metadatos de imágenes

Google ofrece metadatos para cada imagen en el estándar XMP (el esquema EXIF para propiedades TIFF). Si el formato es TIFF, los metadatos XMP se incluirán también como etiquetas TIFF

Ficheros asociados

a) Fichero *checksum*. Un fichero "*checksum.md5"* contiene el *checksum* de cada página del libro para garantizar su autenticidad.

b) Fichero de OCR en UTF-8: un fichero para cada página, con el mismo nombre de la página y la extensión *.txt*.

c) Cada directorio de imágenes lleva un fichero *pagedata.html* con información de las imágenes del volumen (formato, calibración, paginación, información de etiquetas de página) detectados durante el procesamiento automático por Google.

9. CALIDAD

Google efectúa dos controles de calidad. Uno, empleando un algoritmo para detectar errores y otro de carácter manual. Los principales tipos de errores son:

1. texto desenfocado
2. texto oscuro
3. texto descolorido o sobreexpuesto
4. texto torcido
5. texto curvado
6. texto cortado
7. errores en la coloración de la página

Los errores se valoran como críticos, no críticos y cosméticos, según el impacto que tenga sobre la lectura del contenido.

10. INFORMACIÓN A LOS SOCIOS

Para cada volumen escaneado Google ofrece la siguiente información en su interfaz GRIN:

- Si el volumen ha sido escaneado.
- Si se han recibido los metadatos de cada volumen.
- Si el ejemplar ha sido añadido al índice de GBS.
- Si el ejemplar contiene algún material que no se ha escaneado.

APÉNDICE A

Google necesita que la Biblioteca le proporcione la siguiente información sobre el identificador (código de barras).

1. Número de sistemas de código de barras que usa la Biblioteca

2. Formato del código de barras:
    - Especificación completa: simbología (CODABAR, CODE ·), CODE 128, etc.); número de dígitos del código de barras.
    - Campos del código de barras: posiciones de dígitos reservadas para el campo del prefijo de la biblioteca, posiciones reservadas para el campo identificador del libro, posiciones reservadas para el campo del dígito de control, una descripción de cualquier campo no mencionado anteriormente.
    - Método de computación del dígito de control.
    - 10 ejemplos del código de barras con los campos descodificados según lo mencionado anteriormente.

3. Información sobre la unicidad del código de barras:
    - Sólo puede usarse un código de barras por ejemplar. Explicar qué hacer en caso de existir más de un código de barras.
    - Probabilidad de duplicación de un código de barras. Explicar situaciones donde podría darse esta circunstancia.
    - Protocolo de código de barras para revistas y obras en varios volúmenes.
    - Otros protocolos de códigos de barras no mencionados anteriormente.
    - Descripción de métodos para evitar duplicaciones de códigos de barras. Cómo se generan los códigos de barras (a través de una empresa o localmente). Nombre de la empresa que los genera. En caso de generación local, describir el software utilizado y probabilidad de error en la generación de los códigos de barras.
    - Describir el procedimiento a seguir en caso de existir varios códigos de barras en un volumen: ¿enlazan al mismo registro bibliográfico?, ¿se listan todos los códigos de barras en el registro?, ¿qué código de barras es el que se lista en el registro?
    - Qué campos contienen el código de barras en el registro MARC y en qué forma (lista separada por comas, múltiples etiquetas, etc.)
    - Cómo y dónde está situada otra información a nivel de ejemplar en el registro MARC (número de volumen, signatura…)
    - Información sobre la cobertura del código de barras:
        - ¿Cuántos libros tienen código de barras?
        - ¿Cuáles son los planes para poner a los libros que no los tienen?
    - Herramientas asociadas a la lectura de códigos de barras (pistolas, lápices…):
        - Comunicar el fabricante y modelo de las herramientas utilizadas habitualmente.
        - ¿Existen algunas dificultades de lectura con estas herramientas?. En caso afirmativo, comentar la frecuencia de los problemas
    - Otra información sobre el procedimiento del código de barras.

Apéndice B. GRIN

Apéndice C. Información de etiquetado de páginas para pagedata.html y ficheros en volúmenes pagedata.txt