Grupo de Trabajo BUC-Google

Documentos BUC

Serie: Informes

Reunión de socios de Google Book Search - Biblioteca

26 -27 de julio de 2007

Universidad de Michigan. Ann Arbor, EEUU

17 de agosto de 2007

Preparado por: Manuela Palafox (Servicio de Edición Digital y Web)
Isabel Costales (Directora de la Biblioteca de la Facultad de Derecho)

Google Book Search. Library Summit

Programa

25 de julio:

- Visita al Digital Commons
- Asistentes:
  - Biblioteca Nacional de Baviera
  - Universidad de Gante
  - Universidad de Madrid

26 de julio:

- Visita al Faculty Exploratory y Knowledge Navigation Center
- Asistentes:
  - Biblioteca Nacional de Baviera
  - Universidad de Gante
  - Universidad de Madrid
- New Partner Orientation. Orden del día de la reunión: novedades sobre Google Book Search, operaciones y temas relacionados con los procesos de ingeniería y control de calidad.
- Asistentes:
  - Biblioteca Nacional de Baviera
  - Committee on Institutional Cooperation
  - Universidad de Gante
  - Universidad de Harvard
  - Universidad de Madrid
  - Universidad de Princeton
  - Universidad de Wisconsin - Madison
- Visita al Centro de escaneado. Michigan ISC
- Asistentes:
  - Biblioteca Nacional de Baviera
  - Committee on Institutional Cooperation
  - The New York Public Library
  - Universidad de Gante
  - Universidad de Harvard
  - Universidad de Madrid
  - Universidad de Michigan
  - Universidad de Oxford
  - Universidad de Princeton
  - Universidad de Texas - Austin
  - Universidad de Virginia
  - Universidad de Wisconsin – Madison

27 de julio:

- Library Summit. Orden del día de la reunión:
  - Presentación de las funcionalidades de Google Book Search
  - Google Book Search y las editoriales. Información sobre las editoriales europeas
  - Temas legales. Los derechos de autor
  - Ingeniería. Control de calidad
  - Discusión sobre el formato de las reuniones de socios
- Asistentes:
  - Biblioteca Nacional de Baviera
  - Committee on Institutional Cooperation
  - The New York Public Library
  - Universidad de Gante
  - Universidad de Harvard
  - Universidad de Madrid
  - Universidad de Michigan
  - Universidad de Oxford
  - Universidad de Princeton

- Universidad de Texas - Austin
- Universidad de Virginia
- Universidad de Wisconsin – Madison

Aspectos centrales de la reunión

Descripción general de Google Book Search (GBS):

- Adam Smith señaló que los contenidos de GBS se han multiplicado por seis desde enero de 2006 hasta enero de 2007.
- Actualmente, están participando en el proyecto GBS 26 bibliotecas (Ben Bunnell anunció la incorporación de la Keio University, Tokio, Japón) y 10.000 editoriales.
- GBS está disponible en 10 idiomas.
- Los libros de GBS proceden de 100 países. La empresa sigue avanzando en la expansión internacional; quieren entrar en contacto con más bibliotecas y más editoriales. El objetivo es digitalizar el mayor número posible de libros, teniendo en cuenta que, según cálculos de Google, existe un 20% de libros en el dominio público y un 5% de libros en venta y, por tanto, hay que conseguir digitalizar el resto.
- Es esencial mantener la confidencialidad del producto en relación a los siguientes aspectos:
    - Número de libros digitalizados en las bibliotecas
    - Actividades desarrolladas en todos los procesos implicados en la digitalización
    - Lugares donde están situados los centros de escaneado
    - Nombres de los responsables de Google
- La empresa tiene un objetivo muy claro en cuanto a la necesidad de digitalizar libros en otros idiomas distintos al inglés
- Van a crear una nueva página principal para GBS en la que se van a incluir grandes categorías por las que el usuario podrá navegar. Están haciendo análisis de colecciones.
- Hay proyectos de cooperación con OCLC relacionados con derechos de autor, control de autoridades...
- En la reunión también se trató cómo son las relaciones de Google y los socios con la prensa y con otras bibliotecas que no están en el proyecto. Propusieron a las bibliotecas participantes en GBS que si tenían mucha presión de los medios se pusieran en contacto con Google para pedir consejo. Por otra parte, nos invitaron a los socios a organizar conferencias con la comunidad universitaria para explicar el proyecto GBS.

Búsqueda

- En cuanto a las búsquedas, Google quiere mejorar las funcionalidades. Los usuarios podrán hacer consultas por texto completo y metadatos: preguntas específicas por títulos de libros, autores, preguntas temáticas y obtener resultados agrupados por materias y también preguntas de tipo informativo. Se pretende obtener un mayor refinamiento en la búsqueda. Los metadatos de los libros proceden de distintas fuentes: bibliotecas, editoriales y la Web.
- Quieren incorporar mas funcionalidades características de la Web 2.0. El usuario podrá participar, incluir reseñas y valoración de libros, personalizar su página con sus libros favoritos (My Library). La experiencia de los usuarios puede mejorar las búsquedas en GBS (Social Book Search). Para ello, tendrán que abundar en la segmentación de usuarios mediante técnicas extractivas (minería de datos).
- Google está trabajando en profundizar en las conexiones y enlaces entre los libros: de qué forma distintos autores han usado una cita en sus obras. Esta interesantísima funcionalidad va a permitir identificar plagios.

Editoriales

- Dos tercios de las editoriales europeas están cooperando en el Programa Google Partners, de Google Book Search.
- Google está estudiando la posibilidad de firmar acuerdos con las editoriales para que los usuarios puedan leer libros sujetos a derechos de autor por un coste de 4,99$.
- Los ingenieros de Google han observado que cuando se incrementa el número de páginas visualizadas, el usuario pincha más en el vínculo “Comprar”.

- En GBS va a aparecer la lista completa de las editoriales implicadas en el proyecto. Hasta ahora algunos autores habían criticado la falta de transparencia al ocultar esta información que está visible en otros buscadores como Scirus (Elsevier) y Live (Microsoft)

Actualización de las operaciones

- Google pretende incrementar la productividad de las actividades de los socios.
- La productividad de los trabajadores encargados del escaneado sigue aumentando. Asimismo, quieren agilizar la devolución de los libros ya escaneados.
- Sólo se hará el segundo escaneado si ha habido problemas en el primero, no por sistema como se ha hecho hasta ahora.
- Se pretende identificar los errores cuanto antes para no tener que escanear dos veces. (Manifest Reconciliation Process)
- Los índices de error continúan descendiendo: actualmente están por debajo del uno por ciento. Calculan que hay un error cada 200 páginas.
- Google propone trabajar de forma más estrecha entre las bibliotecas y Google para no duplicar las obras ya digitalizadas. Hay que hacer un análisis de las colecciones que permita detectar los duplicados y piden mayor participación de los socios en la identificación de colecciones de interés.
- Google tiene un plan de contención del gasto.
- Google no quiere instalar nuevos Centros de digitalización, sino aumentar la productividad de los ya existentes (Montain View, Michigan, Massachussets, Oxford, Madrid, Alemania)
- Hay un plan para escanear más libros, quieren utilizar nuevas estaciones que no dañen los libros en la encuadernación y se puedan escanear los libros de gran tamaño (en Harvard tienen un 35% de libros de fondo histórico que no pueden escanear)
- En nueve meses Google sacará las nuevas cámaras móviles que permitirán escanear los libros de gran tamaño.
- Google propone a los socios incluir nuevos contenidos. Por un lado, incluir en GBS libros ya digitalizados en las bibliotecas (como es el caso de la Biblioteca Digital Dioscórides) y también otros tipos de materiales, como microfilms y microfichas. A Google le interesa cualquier tipo de material en microfilm, por ejemplo, periódicos, material de archivo, manuscritos...

Actualización de las actividades de ingeniería. Control de calidad

- Auditoría semanal de las imágenes. Criterios:
  - Blur
  - Finger
  - Obscure
  - Marking
- Auditoria semanal del resultado final, después de pasar el primer control. Criterios:
  - Blur (borroso)
  - Crop
  - Dewarp (quitar la curvatura)
  - Cleaning (limpiar)
  - Skew (torcido)
- Los libros se escanean en la estación de escaneado. Después se quita la curvatura (dewarp), se procede a limpiar la imagen y se pasa el OCR. Los siguientes procesos son análisis e indexación de las imágenes.
- En el control de calidad se tiene en cuenta la calidad de las imágenes y la secuencia de las páginas.
- Adquisición:
  - Índice de errores
  - Índice neto de errores
- Resultado final:
  - Índice de errores
- Continúa mejorando la calidad de las imágenes. El índice de error neto es menor del 0,5%, incluyendo las páginas desaparecidas.
- En el departamento de ingeniería de Google Book Search están trabajando en el desarrollo de un software de protección de las imágenes borrosas.

- Actualmente, el 60% de los libros se escanea una sola vez y el 40% dos veces. La decisión de escanear la segunda vez se debe hacer en un plazo de 24 horas.
- El orden de las páginas todavía tiene problemas significativos: entre un 40% y un 50% de los libros les falta al menos una hoja.
- En el primer semestre de 2008 se van a reprocesar todas las imágenes contenidas en GBS.
- Actualización de la aplicación GRIN. Cuando esté disponible una versión mejorada de las imágenes lo notificarán a las bibliotecas.

Aspectos legales

- Actualmente, Google tiene tres querellas judiciales (una en Alemania)os problemas legales más relevantes se centran en las "obras huérfanas" y en el asunto del "fair use".
- **Un asunto muy interesante** en el que Google está trabajando, y en el que tienen mucho interés en que colaboren las bibliotecas participantes en GBS, es el estudio de las bibliografías de los autores y las fechas de fallecimiento de los mismos, para poder determinar las obras que son de dominio público. Para ello, Google quiere crear una base de datos que podamos mantenerla entre todos los socios, con información sobre legislación, autores... La Universidad de Michigan preguntó cuándo estaría disponible esta aplicación y Google respondió que todavía no lo sabían, pero que estaban iniciando un proyecto de cooperación con OCLC sobre estos temas. Recuerdo que hace meses, Ben Bunnell nos preguntó cuál era nuestra forma de proceder a la hora de incluir información en el campo de autor sobre fechas de nacimiento y muerte. Le contestamos que se trataba de un trabajo muy tedioso y lento y que los bibliotecarios de Normalización rastreaban y buscaban esa información en las autoridades de distintas bibliotecas, la BNA, la Library of Congress, etc.

Propuestas sobre el formato de las reunión de socios

- Google pidió la opinión de los socios sobre el tipo de reunión de socios y la periodicidad de las mismas. Se hicieron las siguientes propuestas:
  - Mantener el formato actual, dos reuniones al año.
  - Hacer una única reunión al año.
  - La Biblioteca Nacional de Baviera y la Universidad de Gante propusieron una reunión en Europa a la que asistirían los socios europeos. Nosotras no nos adherimos a la propuesta. Yo no tengo claro la necesidad de reuniones específicas de socios europeos y no se me ocurren qué ventajas podríamos obtener de las mismas.