Encuentro Google Libros. Programa Bibliotecas Octubre 2009

Campus Google. Mountain View, CA

Zacarías Martín Maté
Manuela Palafox Parejo
Universidad Complutense de Madrid

Índice

- Reunión del *Pre-Summit (20 de octubre de 2009)*
- *Summit (21-22 de octubre de 2009)*
  - Google Libros y la evolución de la búsqueda. Erez Lieberman & J. B. Michel
  - Acceso Institucional
  - Características de las novedades de Google Libros
  - Google Ediciones
  - Derechos de autor
  - Calidad de las imágenes
  - Lengua, traducción y OCR
  - Corpus de investigación
  - Reunión con los socios internacionales

PRE-SUMMIT

Universidad Complutense Madrid

Pre-Summit
Universidad de Stanford

- Las bibliotecas dieron datos sobre:
  - Las colecciones digitalizadas.
  - Número de ejemplares a digitalizar y digitalizados. Producción diaria. Ejemplos: California digitaliza 3.000 libros al día; Michigan tiene ya digitalizados mas de 4 millones y digitaliza 2.500 diarios; Keio digitaliza 300 al día; Princeton tiene digitalizados 700.000 y tendrán al final del proyecto un millón; Columbia digitalizará 500.000 y hasta ahora han digitalizado 50.000. En Lyon tienen 200.000 libros en el dominio público…
  - Herramientas empleadas (selección con lista o selección directa desde la estantería).
  - Almacenamiento de la copia digital y Plan de preservación a largo plazo.
  - HathiTrust. A shared digital repository: servicios de almacenamiento, preservación y datos bibliográficos
  - Participación en el nuevo acuerdo (Settlement). Estatus de participación. Suscripción institucional. Debería haber contactos entre los abogados de las distintas instituciones. Texas no firma el acuerdo y presentó un texto explicando su posición.
  - Prioridades, desafíos y preocupaciones:
    - La calidad de las imágenes y la métrica de la calidad (métrica automática y manual). Implicar a los usuarios en la detección de los errores.
    - Los duplicados, distintas ediciones. Google no debería rechazar los libros de distintos editores y/o distintas fechas de publicación.
    - Los metadatos de facticias y chino.
    - ¿Por qué no incluyen todos los metadatos en Google Books?
    - Títulos de revistas, series.
    - GRIN: ¿cada cuánto tiempo vuelven a descargar las imágenes? Proponen incluir la fecha de descarga de las imágenes mejoradas.
    - Sobre la investigación del corpus hay varias universidades interesadas en trabajar en ello. Muy pocas podrán hacerlo porque la inversión es muy grande (alrededor de 5 millones de $). Tendrán que reunirse y decidir qué servicios van a proporcionar.

Encuentro Google Libros: Programa Bibliotecas. Campus Google, Mountain View, CA.
Octubre, 2009

Universidad Complutense Madrid

2009 LIBRARIES SUMMIT

Resumen general de las novedades
Dan Clancy

- Mejoras en los OCR. Detectar imágenes, ilustraciones, leyendas
- Nuevo formato: epub para descargar los libros de dominio publico
- Versión Web de móviles.
- Nuevos Socios: B&N y Sony entre otros.
- Opción de Licencias Creative Commons para las editoriales.
- Contenido de Google Libros: 12 millones de libros. Alrededor de un millón de duplicados.
- Cada día se usan el 40% de los libros de las bibliotecas, 17% de libros de dominio público.
- Cada mes, los usuarios pre-visualizan el 81% de los libros de las bibliotecas, un 21% de libros de dominio público.
- API's de Google Books
- Mercado del e-book. Su intención en **Google Editions** es que el usuario pueda comprar desde cualquier dispositivo: E-book, móvil, notebook, portátil…

Encuentro Google Libros: Programa Bibliotecas. Campus Google, Mountain View, CA.
Octubre, 2009

Universidad Complutense Madrid

Actualización del corpus digitalizado
Kart Groetsch

- El 85% de los libros de Google Books proceden de las bibliotecas (11 millones) y el 15% de las editoriales (2 millones):
  - Oxford 41%,
  - Baviera 24%
  - Gante 10,88%,
  - Lausanne 10%
  - Madrid 7,38 %,
  - Cataluña 4,30% y
  - Keio 1,5%
- Michigan el 34%, 17% California, 16% Stanford, Harvard, 7,98% Texas, 4,47, Virginia 3,55%, Wisconsin 2,70. San Diego 2,37, New York, 2,23% , etc.

Actualización del corpus digitalizado
Kart Groetsch

- Fondos en 478 lenguas; el 47% de lenguas no inglesas: alemán 9,59%, 6,5% francés, 6,04% español.
- País de edición: 34,8% de UK; 12,91% Alemania; 8,57% Francia; 5,84% Suiza 3,16%; Rusia 2,58; India 2,48%; Canadá 2,46; Italia 2,46%; Japón 2,42%; España 1,75%. En total: Europa el 40% y USA el 39%.
- Accedidos de la semana 4 al 10 de octubre (un 17,96% del total)
  - NYPL: 37,39%
  - UCM: 33,15%
  - Oxford: 30,40%
  - Harvard: 29,71
  - Cataluña: 28,15%
- Tráfico de países: USA 49,20%. Francia 8,53%, Italia 5,24%, Alemania 4,89%, Canadá 3,23%, UK 3,10%, SP 2,87%, México 1,25%, India 1,24%, Brasil 1,22%, Bélgica 1,18%, Australia 1,10% y Argentina 0,99%

Encuentro Google Libros: Programa Bibliotecas. Campus Google, Mountain View, CA. Octubre, 2009

Universidad Complutense Madrid

Google Libros y la evolución de la búsqueda

Erez Lieberman & J. B. Michel

- En este proyecto utilizaron los contenidos de Google Books, pero los investigadores no tuvieron acceso a los libros por problemas de derechos de autor; Google les dio solamente datos estadísticos.
- Entre los distintos análisis que hicieron, estudiaron las frecuencias de verbos regulares e irregulares en lengua inglesa. Con el uso, los verbos se regularizan. Si un verbo es 100 veces menos frecuente, se regulariza 10 veces mas rápido.
  - Ejemplos:
    - thrived/throve. El gráfico del uso de este verbo desde 1800 hasta ahora explica la tendencia.
    - Ejemplos de gráficos con las dos formas de cada verbo: líneas de la forma regular e irregular. Batalla de las terminaciones *t y ed*:
      past **strove** *or* **strived**; sped vs speeded

Google Libros y la evolución de la búsqueda

Erez Lieberman & J. B. Michel

- Otras aplicaciones:
  - Lexicografía:
    - Gráfico de uso de nuevas palabras añadidas al diccionario: ejemplo: *albuterol* (en el gráfico asciende en 1950 y luego baja)
  - Historia geopolítica:
    - Cambios en los nombres de los países: Ghana, Rodesia, Zimbawe, Malawi…
  - Historia de la ideología. Terminación en "*ism*"
    - Comunism / Capitalism / Socialism / Terrorism
  - Epidemiología histórica:
    - 1890. Gripe rusa
    - 1930. Gripe española….
    - Por ejemplo el cólera. Cómo aparece este tema en los libros durante los momentos críticos de las epidemias.
    - Historia de la medicina. Descubrimiento de Fleming de la penicilina en 1928, producción masiva en 1943.

Encuentro Google Libros: Programa Bibliotecas. Campus Google, Mountain View, CA.
Octubre, 2009

Universidad Complutense Madrid

Google Book Search

- Posibilidad de meter Google Libros en una página Web mediante un objeto embebido, no me queda claro si un iframe o un embed, estilo YouTube.
- Van a poner nuevas opciones, como que es nuevo, el top de libros, mi biblioteca, con opciones como: Estoy leyendo, para leer, leído, comprar.
- En sección estoy leyendo con favoritos, organizar, compartir.
- Nubes de etiquetas.
- EPUB formato para ebook.
- Sacarán otras ediciones y libros relacionados. Palabras y frases comunes, lugares mencionados en el libro.
- Se podrá ver dos paginas a la vez, con flecha derecha e izquierda para pasar paginas.
- Interfaz para móviles, con los 3 mas recientes leídos.

Calidad

- Porcentaje de libros imperfectos, no incluidos el 5% de libros con paginas missing, ha evolucionado de la siguiente forma: Jun 2007 30%, Feb 2008 20%, Sept 2009 12%.
- El chequeo automático funciona razonablemente bien comparando 2 versiones de la misma pagina, pero no es buena para una pagina dada.
- Así que les gustaría reemplazar los chequeos automatizados por otro tipo de chequeo basado en la premisa de que la mayoría de los libros suelen ser o muy buenos o muy malo. Se van a tomar meses para hacer esto en todos los trabajos que ya han escaneado.
- Problemas sistemáticos en un 5% de los volúmenes.
- Reciben 700 informes de usuarios por día.
- Páginas missing en el 13% de los volúmenes.
- En una posible suscripción, te vendemos el libro a menor precio cuando tiene defectos, o que se hace.

Google Ediciones

Abe Murray, Dan Clancy, et al.

- Este nuevo servicio estará operativo en la primavera de 2010. El objetivo es llevar miles de títulos a cualquier dispositivo sin importar los formatos de las obras, ni la marca del dispositivo. El objetivo de Google es que el usuario pueda comprar y leer los libros desde cualquier sitio. El usuario busca en Google, encuentra un libro y puede pagar vía cuenta de Google o a través de un minorista.
- Editoriales. 2 millones de libros. 30.000 editoriales
- ¿Donde están los clientes? Hoy el mercado del e-book está fragmentado: Amazon, B&N, e-reader de Sony, etc.

Idioma, Traducción y OCR

Franz Och and Ray Smith

- Traducción automática. La tecnología puede romper la barrera del lenguaje. Se están utilizando los millones de libros de Google Books y se ha avanzado mucho en la calidad de la traducción de la mayor parte de las lenguas, aunque todavía tienen mucho trabajo por hacer.
- La traducción automática tiene problemas para traducir algunas lenguas, por ejemplo el hindi al inglés.
- Mejora en la calidad debida a las siguientes razones:
  - Enfoque estadístico / aprendizaje
  - Modelos de traducción más sofisticados
  - Más datos
  - Mayor número de grupos de investigación, interés creciente

Idioma, traducción y OCR

Franz Och and Ray Smith

- El enfoque estadístico. Han tenido que tomar determinadas decisiones en condiciones de incertidumbre. Por ejemplo, un verbo puede tener diferentes significados en el contexto de una frase. El enfoque estadístico nos permite manejar y combinar muchas ambigüedades, combinando grandes cantidades de textos para integrar todo ese conocimiento previo.
- OCR
  - Problemas con las letras itálicas.
  - Múltiples fuentes: Engine OCR comercial, correctores humanos con ReCaptcha.
  - Dificultades de los OCR multilingüe: detección de la orientación de los distintos alfabetos. Líneas horizontales o verticales, izquierda, ambigüedad del tamaño.

Investigación del corpus

Jon Orwant & Abe Murray

- Análisis de imágenes y extracción de textos. Análisis computacional para mejorar la imagen o extraer información textual o estructural de la imagen (por ejemplo OCR). Ejemplos de uso: extracción de leyendas de PubMed.
- Análisis textual y extracción de información. Desarrollo de concordancias, extracción de citas, clasificación automatizada, extracción de entidades y proceso de lenguaje natural. N-grams
- Análisis lingüístico, entender el lenguaje, el uso lingüístico, semántica.
- Traducción automática. Google tiene contenidos en 478 lenguas.
- Logística: transferir datos, actualizar datos, correr programas sobre los datos (Middleware. Proyecto Bamboo), compartir resultados.

¿Cuál es la situación entre los socios internacionales?

Ben Bunnell & Clara Armand-Delille

- ¿Podemos cambiar la fecha de 1869? Baviera sugiere ampliar la fecha hasta 1872. Google dice que no es posible.
- UCM sugiere digitalizar libros bajo copyright. A Ben Bunnell le parece una buena idea, pero habría que tratarlo con las editoriales.
- Paisaje legal / fair use. ¿Cuáles son las perspectivas de las bibliotecas?
- Planes de Google:
  - Comisión Europea- reforma de las leyes de PI- Google editions. Frankfurt Book Fair Merkel. EC Hearing- BNF and Italian Ministry of Culture talks.
  - Regalos a las bibliotecas nacionales. Libros en dominio público a las instituciones culturales en Europa. Compartir libros en el dominio público, socios existentes.

Accesos a los libros Complutenses

Período de 7 días: 4-10 de octubre de 2009

Accesos	Título	Autor	Año
12.490	Diccionario etimológico de la lengua castellana (ensayo)	Pedro Felipe Monlau	1856
12.008	Diccionario geográfico-estadístico de España y sus posesiones de ultramar	Pascual Madoz	1830
8.637	La Ilíada	Homero	1788
8.275	Vida y viajes de Cristóbal Colón	Washington Irving	1852
7.520	Enciclopedia moderna	Francisco de Paula Mellado	1851
7.027	Los tres reinos de la naturaleza o museo pintoresco de historia naturaleza: Botánica. Mineralogía	Georges-Luis Leclerc Buffon	1858
6.468	Diccionario de la lengua castellana	Real Academia Española	1852
4.179	Diccionario de agricultura práctica y economía rural	Agustín Esteban Collantes, Agustín Alfaro	1855
4035	Anatomie descriptive	Jean Cruveilhier	1837

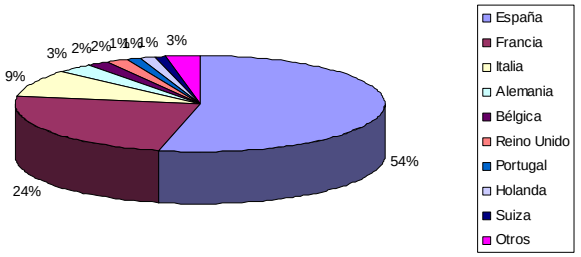
Encuentro Google Libros: Programa Bibliotecas. Campus Google, Mountain View, CA. Octubre, 2009

Universidad Complutense Madrid

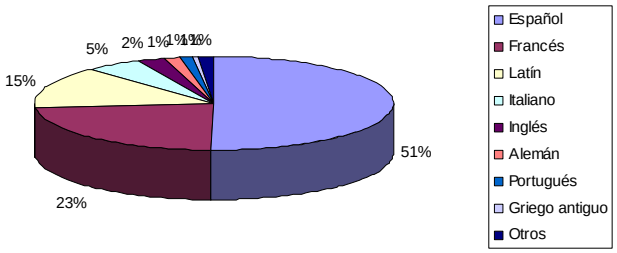
Accesos a los libros Complutenses

Período de 7 días: 4-10 de octubre de 2009

País de edición



Idioma

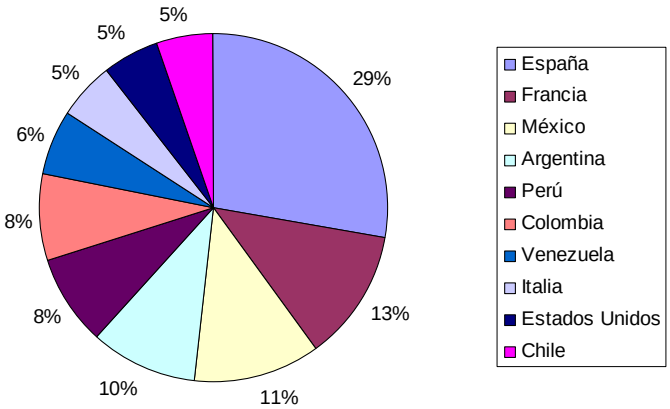


Universidad Complutense Madrid

Accesos a los libros Complutenses

Período de 7 días: 4-10 de octubre de 2009

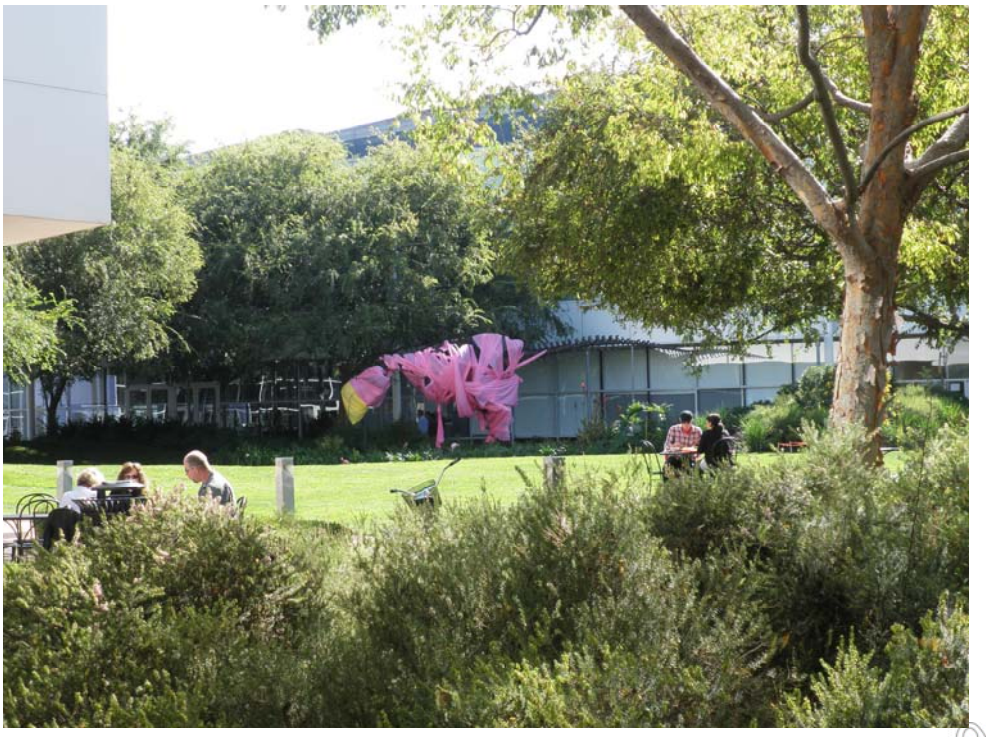
Tráfico por país



Encuentro Google Libros: Programa Bibliotecas. Campus Google, Mountain View, CA.
Octubre, 2009

Universidad Complutense Madrid

Sede de Google
Mountain View, CA



Universidad Complutense Madrid

Sede de Google
Mountain View, CA

Encuentro Google Libros: Programa Bibliotecas. Campus Google, Mountain View, CA.
Octubre, 2009

Universidad Complutense Madrid

Sede de Google
Mountain View, CA

Encuentro Google Libros: Programa Bibliotecas. Campus Google, Mountain View, CA.
Octubre, 2009

Universidad Complutense Madrid

Sede de Google
Mountain View, CA

Encuentro Google Libros: Programa Bibliotecas. Campus Google, Mountain View, CA.
Octubre, 2009

Universidad Complutense Madrid

Sede de Google
Mountain View, CA

Universidad Complutense Madrid

Sede de Google
Mountain View, CA

Encuentro Google Libros: Programa Bibliotecas. Campus Google, Mountain View, CA.
Octubre, 2009

Universidad Complutense Madrid