# EVENTOS

V WORKSHOP REBIUN PROYECTOS DIGITALES. LA BIBLIOTECA DIGITAL Y EL ACCESO A NUEVOS CONTENIDOS

2005

## DIGITALIZACIÓN MASIVA Y ACCESO EN LÍNEA A PRENSA REGIONAL

crue Universidades Españolas
Red de Bibliotecas REBIUN

Universidad de
Las Palmas de Gran Canaria

BULPGC. Servicio de Biblioteca y Documentación

# Digitalizacion masiva y acceso en linea a prensa regional

http://bdigital.ulpgc.es/prensa

Víctor M. Macías Alemán y Javier Torres Betancor

Email
sdd@ulpgc.es

V Workshop de proyectos digitales REBIUN. Barcelona, 2005

# Inicios

- Comienzo abril 1996
- Incluir toda la prensa corriente
- Solucionar:
  - Almacenamiento
  - Preservacion
  - Completar colecciones
  - Acceso unico
- Facilitar consultas (futuras) localmente

# Inicios

- Participacion abierta: bibliotecas publicas y privadas, municipales y universitarias, gabinetes de prensa, etc.
- Alguna microfilmacion previa
- Contratacion empresa
- Presupuestos ordinarios

# Procedimiento

- Envio periodicos a empresa
- Formato TIFF 1bit ITU-TT.6
- Almacenamiento en CD-R estandar
- Visor monopuesto
- OCR
  - txt
  - mdb
- Recepcion y control de calidad
- Acceso publico localmente

# Inconvenientes

- Demora recepcion digitalizaciones
- Errores grabacion, omisiones
- Presupuestos solo publicos
- No cubre retrospectiva
- Acceso local, horario limitado

# Busqueda de soluciones

- Cambio empresa
- Participacion editoriales prensa
  - Mas calidad
  - Totalidad cabecera retrospectiva
  - Disponible para la venta
- Cambio formatos y sintaxis
- Servidor biblioteca digital BULPGC (acceso Intranet)

# Estudio acceso en linea

- Carga a servidor de 2,8 millones paginas en 2300 CD/DVD
- Aumento capacidad servidor
- Plataforma software libre LAMP
- Unificacion de acceso a formatos y sintaxis diversa
- Utilizacion de herramientas y procedimientos estandar

# Cambio estructura y formato
(enero 2004)

- DVD-R estandar 4,7 Gb
- PDF multipagina JBIG-2,
  OCR imagen orig. texto oculto,
  300 ppp
- Unico fichero periodico-dia
- Sintaxis ficheros autoidentificable
- Sin estructura directorios

# Planificacion

- Aumento capacidad servidor
  200 Gb – 4 Tb (RAID 5 + mirror)
- Fases:
  - Consulta cronologica imagenes
  - Busqueda textual + imagenes
- Descartada conversion global a PDF
- Carga en servidor FTP seguro
- Programacion PHP + librerias (conversion al vuelo TIFF-PDF)
- Acceso transparente al usuario
- Metadatos DC y WAI-A
- Validacion externa PAPI

## Estado actual

- 3,1 millones paginas en Intranet accesibles para 25000 usuarios

## Prospectiva

- Segunda fase:
  - Busqueda textual
  - txt + mdb a MySQL
  - PDX – ocr a MySQL
- Accesibilidad WAI-AAA
- Adopcion XML, OAI-PMH y Marc21, via AbsysNET y Metalib
- Crecimiento 180.000 pags. anuales

# Digitalizacion propia

- Cabeceras historicas
- Extension asumible
- Formato DjVU
  - Compresion 1000:1
    - Doc. Original 2,5 Gb
      - PDF 155 Mb
      - JFIF (JPEG) 128 Mb
      - DjVU 3 Mb
  - Carga paginas sueltas
  - Conversion libre http://djvulibre.djvuzone.org

# Conclusiones

- Construccion biblioteca digital europea se
- puede hacer tambien de abajo a arriba
  Crear consorcios regionales
  unica posibilidad de hacer
  accesible la documentacion local
- *"No más prototipos, es el momento de*
  *aplicacion a gran escala"*
  Conferencia Internacional sobre digitalizacion del patrimonio cultural europeo (Utrech, octubre 1999)