

# Web archiving in a Web 2.0 world

Edgar Crook

*The Electronic Library*, Vol. 27, n. 5 , 2009

Artículo en el que se analizan los proyectos de archivo web desarrollados por la Biblioteca Nacional de Australia. Se describen 3 metodologías de archivo: 1) Proyecto PANDORA: se trata de un proyecto iniciado en 1996 que archiva una selección de publicaciones web australianas tales como revistas electrónicas, publicaciones gubernamentales y sitios web de cierta importancia cultural. Se desarrolló un sistema propio, PANDAS, que va ya por su 3ª versión, con el objeto de facilitar la recopilación, almacenamiento y acceso público al Archivo; 2) Puesto que las publicaciones web recopiladas en PANDORA sólo suponen una pequeña proporción del dominio australiano, en el año 2005, la Biblioteca Nacional contrató a Internet Archive para realizar un web harvesting anual de todo lo que se pudiera recopilar del dominio .au. Esta recopilación se realiza durante un mes cada año. En comparación con PANDORA, la cantidad de datos almacenada es gigantesca. Por ejemplo, en 2007 se recopilaron 18 TB de datos en un solo mes, mientras que PANDORA había recopilado 2 TB en 11 años. A diferencia de PANDORA (donde se obtuvo el permiso de los editores), este segundo archivo web no se ha podido mostrar públicamente debido a dificultades derivadas de la actual ley de copyright australiana y de la ley de depósito legal; 3) La sección de colecciones asiáticas de la Biblioteca Nacional de Australia, contrató el servicio Archive-It, proporcionado por Internet Archive, para recopilar determinados sitios web publicados en Asia. Archive-It es un servicio que, mediante una suscripción anual, se encarga de la recopilación y alojamiento de una serie de URLs seleccionadas. El artículo analiza también los problemas con los que PANDORA se ha ido

encontrando en relación a los formatos de los documentos. En los comienzos, incluso una simple página web diseñada con frames presentaba problemas a la hora de ser recopilada. Aunque muchos de estos problemas se han solucionado, actualmente existen dificultades con los contenidos multimedia, especialmente con los vídeos. Finalmente, se pone de manifiesto que no se está recopilando la enorme cantidad de contenido producido individualmente a través de las tecnologías web 2.0, aunque se está empezando a trabajar en esta línea, por lo que se han firmado acuerdos con Flickr y se han recibido permisos de archivo por parte de MySpace y YouTube.

Resumen elaborado por : Mayte Blasco Bermejo