TALLER I3B: DOCUMENTACIÓN DIGITAL EN BIODIVERSIDAD

Captura y almacenamiento : Biodiversity Heritage Library

Martes 14-abril 2015

Keri Thompson
Digital Projects Librarian
Smithsonian Libraries

Smithsonian Libraries

BHL
Biodiversity Heritage Library

¿quién es este “Keri“?

- Bibliotecario en Smithsonian Libraries
- administrar la digitalización de Smithsonian Libraries colección - que incluye la digitalización de BHL
- miembro de BHL grupo de asesoramiento técnico “TAG”
- coordinar proyectos de datos digitales de Libraries
- Principal de departamento de servicios web
- (También, mi jefe es director del programa de BHL)

Captura

Objetivo - proporcionar una representación digital fiel del objeto original

- una página por imagen
  - (excepto Field note-books - 2 páginas por imagen)
- ningún corregir de la imagen

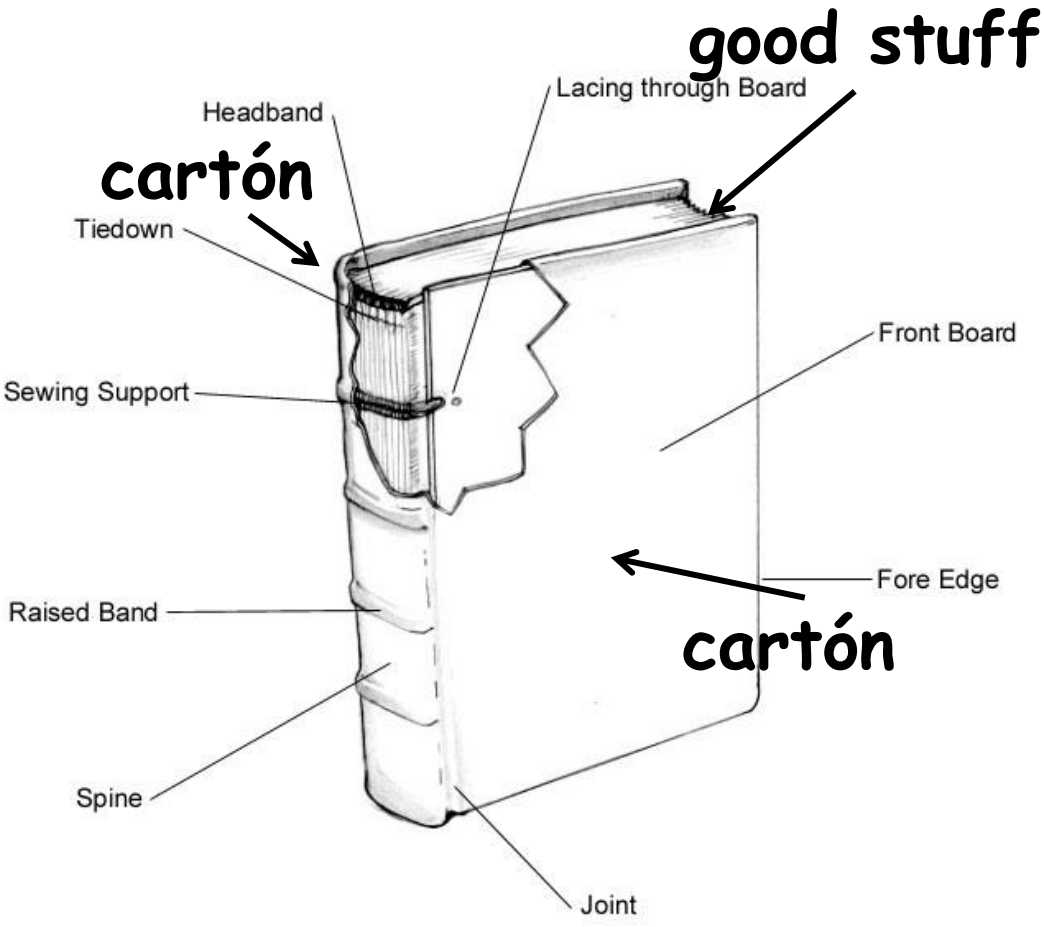
Reutilizar metadatos existentes

- en el catálogo de la biblioteca
- de otras fuentes (BioStor etc.)

- La mayoría de las bibliotecas BHL US/UK utilizan Internet Archive (IA) para la digitalización de libros
  - Open Access, sin fines de lucro
  - Servicios bajo costo
- Los miembros también pueden hacer su propia digitalización, o contratar a otro vendedor
- Los miembros proporcionan metadatos básicos de sus catálogos de bibliotecas

Escanee libros *, de tapa a tapa una imagen por página

* también denominado "volumen" o "item" es un unidad física, no la unidad intelectual, i.e., un libro = múltiples artículos ***o*** un libro= un monográfico



Normas recomendadas

DLF "Benchmark for Faithful Digital Reproductions of Monographs and Serials" http://www.diglib.org/standards/bmarkfin.htm).

Normas preferidos

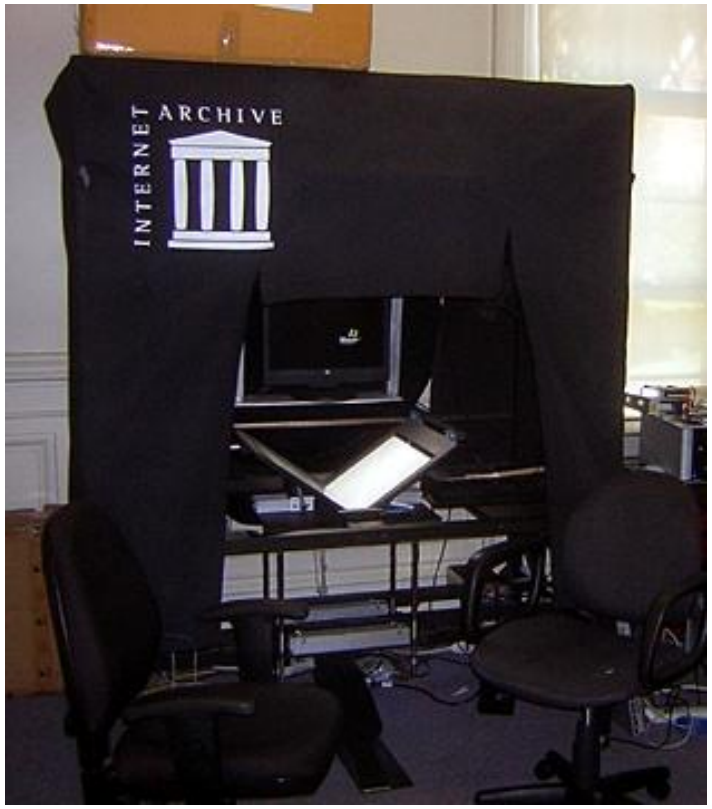
24bit color

> 300ppi

TIFF o

JPEG2000 sin comprimir

Dos caminos principales de captura:



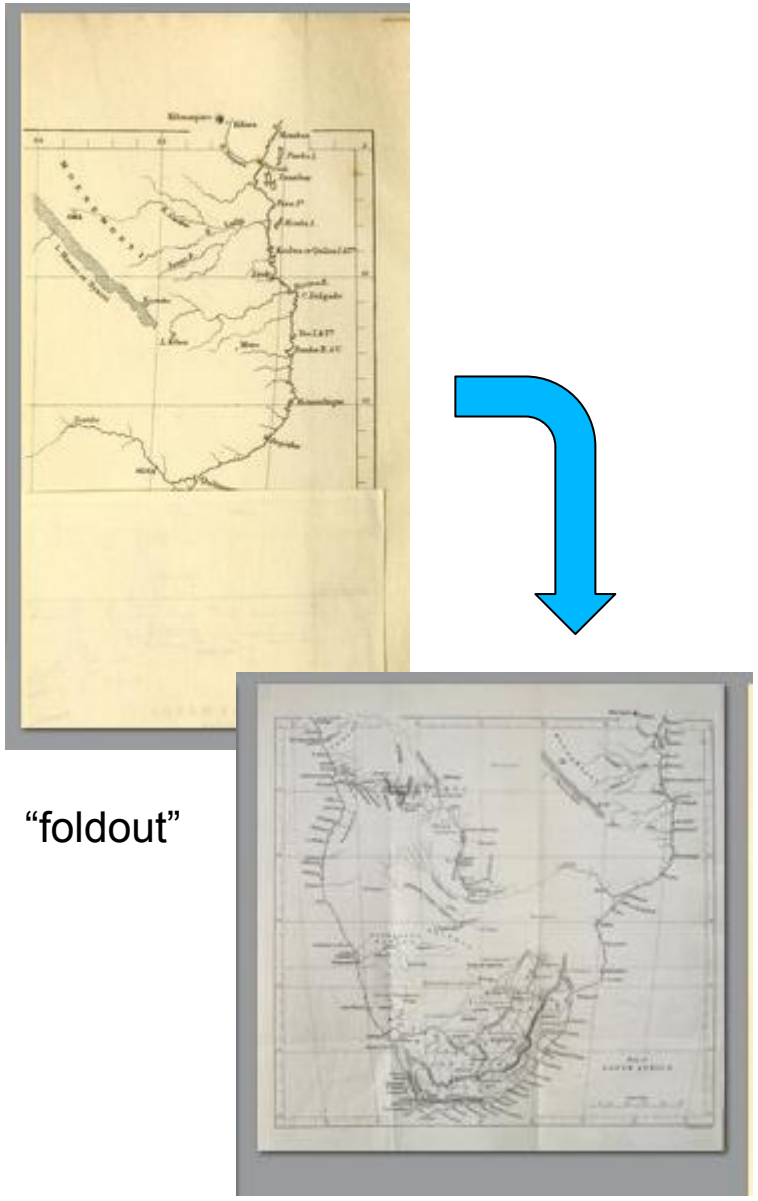
Internet Archive

Digitalización en casa


PhaseOne P65 camera on stand

Internet Archive

Libros digitalizados sur el "SCRIBE"
Metadatos de la biblioteca de Z39.50
Otros metadatos proporciona para IA a través de hojas de cálculo especial "Partner Meta App"
Metadatos a nivel de página entró por el personal IA utilizando su software "biblio"



"foldout"

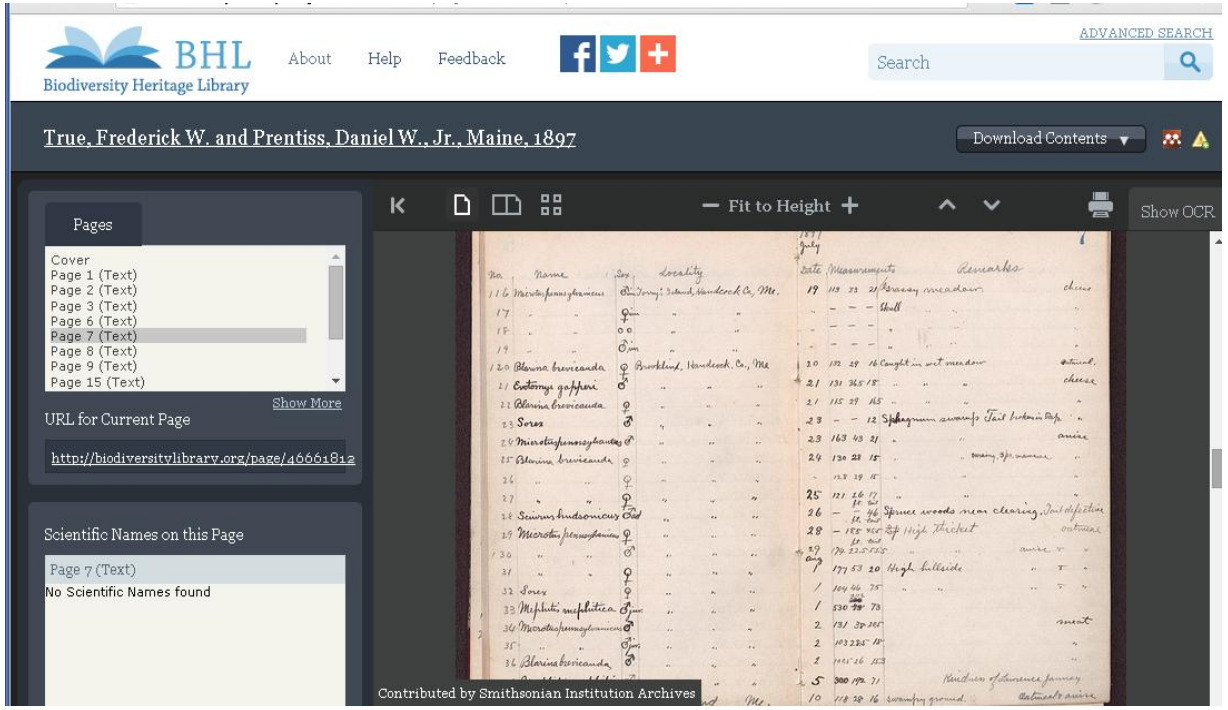
Internet Archive imágenes

Sistema de cámara dual (Canon 5D MkIII)
"Fold-outs" tales como mapas de gran tamaño deben hacerse por separado
Texto que se extiende por la página se debe hacer con las dos páginas en una sola imagen
Resolución varía dependiendo del tamaño del libro
300-600ppi color JPEG2000
Imagens son la compresión JPEG2000 85%, con pérdidas

En casa

Imágenes escaneadas por la biblioteca u otro distribuidor
Metadatos recolectados a través de Z39.50
Los metadatos adicionales para el artículo y páginas entró por personal de la biblioteca utilizando el software Macaw (imita IA biblio software)

Smithsonian Libraries:
utiliza 2 sistemas de Phase One:
P65 60 MP cámara en un soporte de copia y BC100 – dual-cámara 40MP
CaptureOne software
Por folios (> 36cm), libros frágil



EXCEPTO Field Notebooks Proyecto (Smithsonian Archives) - 2 páginas por imagen para notebooks, cartas en escáner de cama plana

Metadata

"Título" registro MARC del catálogo de la biblioteca

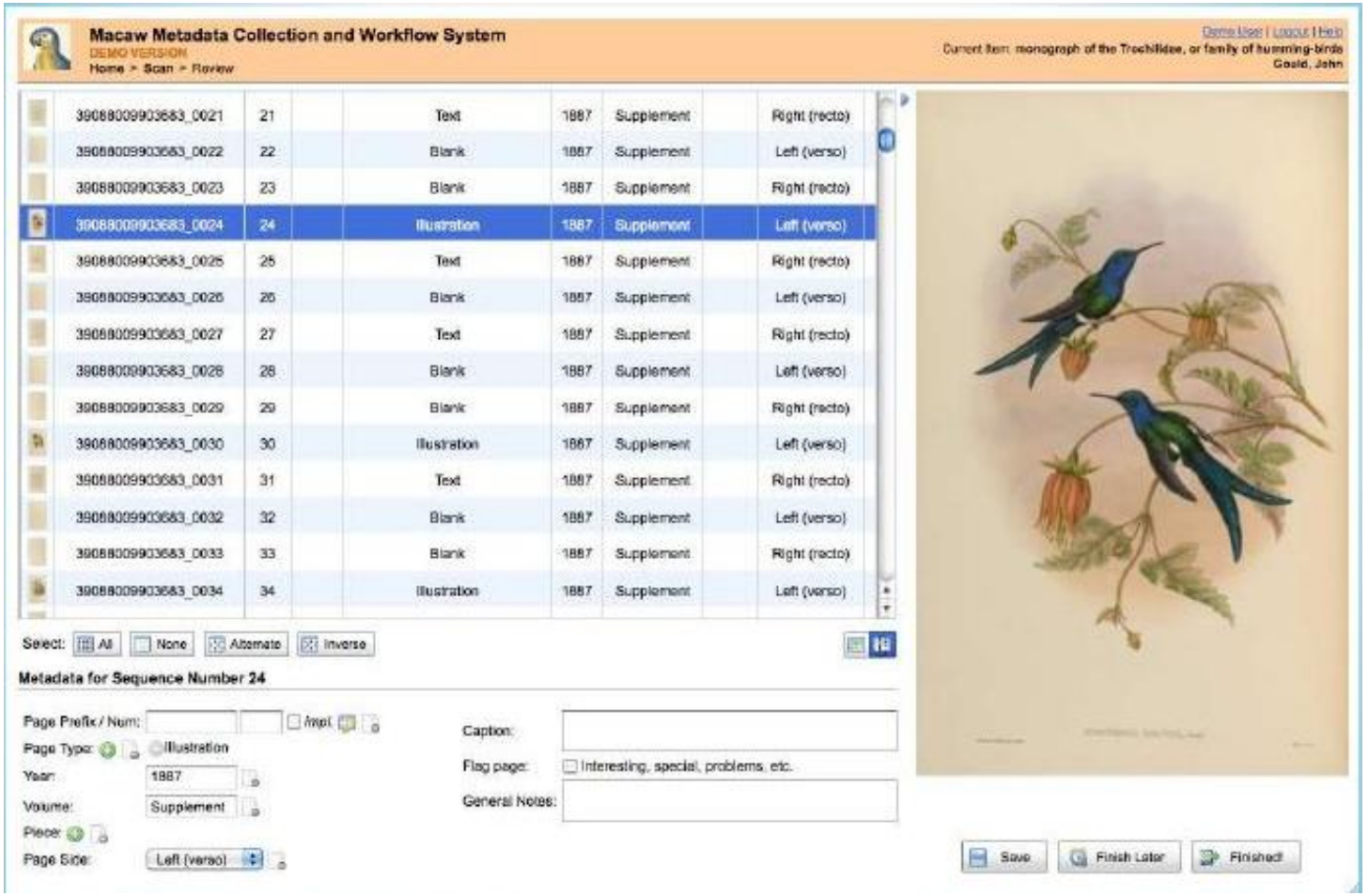
Transformado en MARCXML y MODS

Información "Volumen" de catálogo o introducido por humanos, almacenada en xml

"Segmento" (artículo) la información introducida por humanos o de bioStor etc. (después de la digitalización)

"Página" metadatos entró por humanos, almacenada en el archivo XML que proporciona estructura al objeto digital

Macaw



permite a la gente a entrar en los metadatos a nivel de página como el número de página, tipo de página (ilustración, texto, etc.)

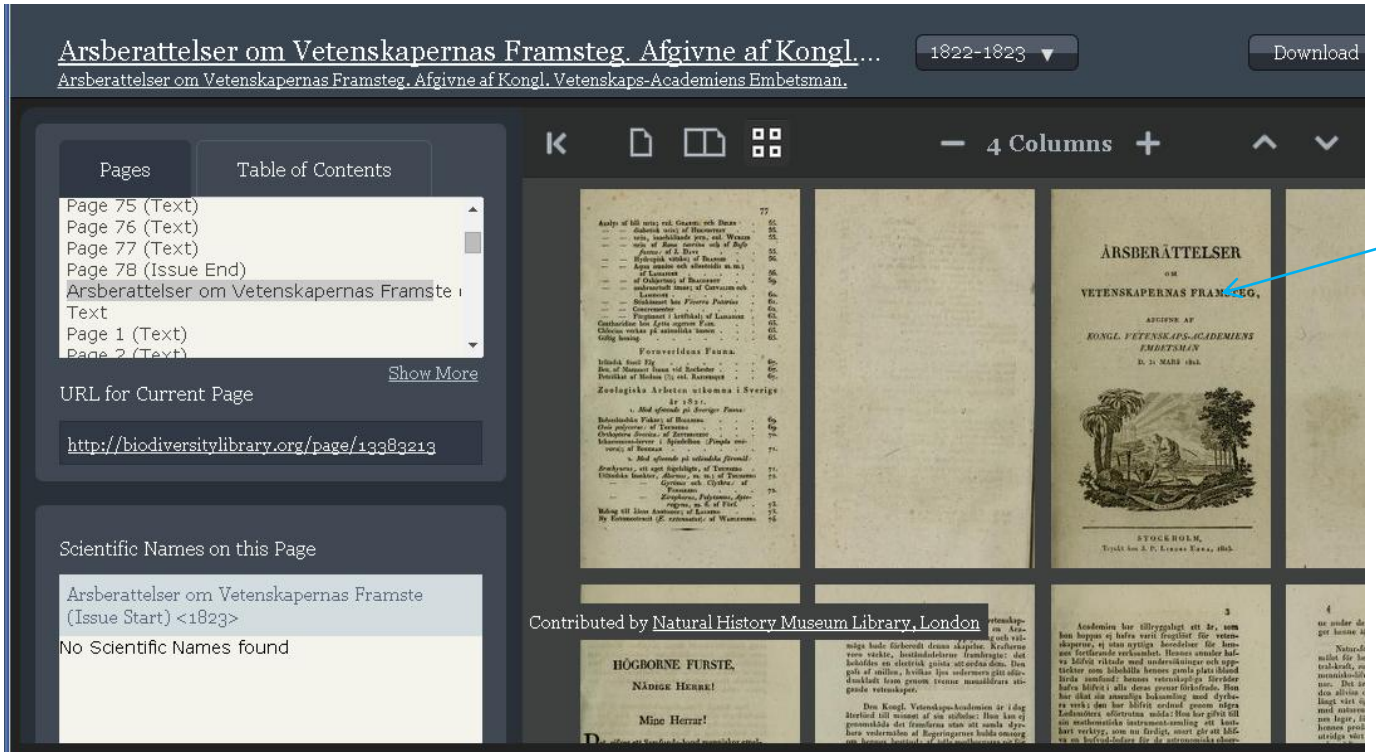
crea archivos XML para subir a IA

Segments (alias articulos, alias citas)

de bibliografías como BioStor, Zookeys (mayoría)
creado por los usuarios (poco)

Titulo →

Titulo de segment



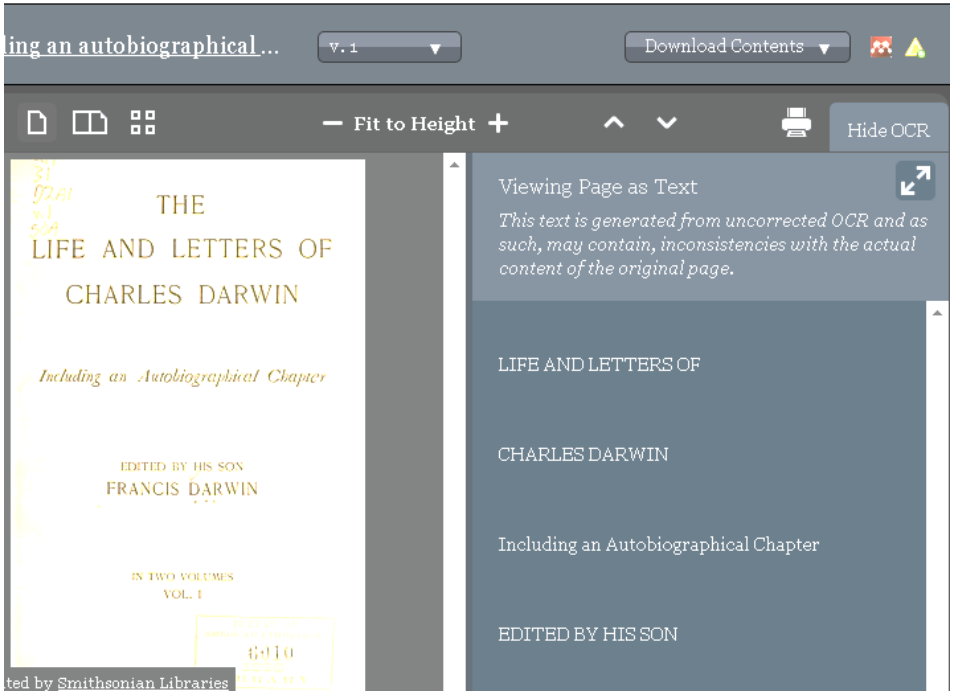
Segment comienza

Otros archivos derivados de originalmente creado por el proceso de Internet Archive

- PDF
- Djvu (OCR text - .txt and .xml)
- ePub/Daisy/Kindle

Otros archivos creados por el proceso de BHL

- Los nombres taxonómicos
- OCR text (diseccionar)
- BHL METS



BHL servidor almacenar

- metadatos descriptivos
- Metadatos para “segments”
- PDFs creados por el usuario (temporal)
- archivo OCR diseccionar
- Los nombres taxonómicos

Internet Archive almacenar

- metadatos descriptivos, administrativos y estructural
- Image files,
- pdfs, ePub etc.
- Djvu (single OCR text file)
- Los nombres taxonómicos
- BHL METS

Index of /2/items/39088001578236/

../		
39088001578236_images_orig/	22-Apr-2005 16:07	-
39088001578236.djvu	22-Apr-2005 19:47	6925667
39088001578236.gif	22-Apr-2005 19:52	411798
39088001578236.pdf	18-Jul-2005 20:17	25181703
39088001578236_Output_processed.zip	22-Apr-2005 16:16	1317611686
39088001578236_dc.xml	09-May-2006 21:34	473
39088001578236_djvu.txt	22-Apr-2005 19:53	629007
39088001578236_djvuxml.xml	22-Apr-2005 19:53	6933514
39088001578236_files.xml	14-Sep-2009 18:55	115123
39088001578236_marc.xml	09-May-2006 21:34	2172
39088001578236_meta.mrc	09-May-2006 21:34	591
39088001578236_meta.xml	14-Sep-2009 18:56	2542
39088001578236_meta_old.xml	09-May-2006 21:34	1397
39088001578236_metasource.xml	09-May-2006 21:34	593
39088001578236_reviews.xml	10-Nov-2008 01:10	1665
BatchProcess.xml	21-Apr-2005 17:21	1271570

Almacenamiento de archivos Primaria y "staging area" está en Internet Archive en San Francisco, USA

Todo es replicado en Alexandria, Egypt

Respaldo secundario está en la Smithsonian, incluye TIFF para los volúmenes escaneados en casa (SIL)
~90TB

Obrigada!