Challenges of Archiving the Web

John C. Gonzalez
Director of Engineering
jcg@archive.org
The Internet Archive: Non-Profit Library
Universal Access to All Knowledge
90,000 Software Titles
90,000 Software Titles
2,000,000 Moving Images
90,000  Software Titles
2,000,000  Moving Images
2,300,000  Book Archive
90,000  Software Titles
2,000,000  Moving Images
2,300,000  Book Archive
2,400,000  Audio Recordings
90,000 Software Titles
2,000,000 Moving Images
2,300,000 Book Archive
2,400,000 Audio Recordings
3,000,000 Hours of Television
90,000  Software Titles
2,000,000  Moving Images
2,300,000  Book Archive
2,400,000  Audio Recordings
3,000,000  Hours of Television
4,000,000  eBooks
90,000
2,000,000 Moving Images
2,300,000 Book Archive
2,400,000 Audio Recordings
3,000,000 Hours of Television
4,000,000 eBooks
505,000,000,000 Web Captures
25,000,000,000,000,000+ Bytes Archived
(Over 25 PetaBytes)

2,500,000 patrons per day
18,000,000 A/V plays per day
How?
Some Principles

- **Transparency**: Items = Directories on Disk
- **Simplicity**: Disk = Unit of storage
- **Preservation**: Each disk is replicated
- **Scale**: BOTH disks serve content
- **Continued Access**: Evolve formats as needed
Challenges

Density of disk drives 2T → 4T → 8T = 8:11:24 @ 2Gbps… practically: 16-17 hours @ 1Gbps; days with interruptions

Complexity of modern content (new web content)

Bad actors

Environment / cost of storage
Thank You!

John C. Gonzalez
Director of Engineering
jcg@archive.org
Details and Back-up
Internet Archive
Item Structure Detail

John Gonzalez • jcg@archive.org • September, 2016
(derived from presentation of Jim Nelson • 11 April 2016)
Overview

What is an item?
Item identifiers
Details page
Item directory
What is an item?

Items are the building blocks of the Archive

An organizational unit: content, collections, user info, & more

A book, a film clip, an album or a song … all could be items

Collections are items—items which “hold” other items

Each item has a unique identifier, e.g.:

coverartarchive
lincolnndouglasde00link
mma_interior_of_a_german_battleship_73634
# Item identifiers

**Strong enough for a human...**

- Often incorporate names and numbers to help identify content

- Partners may add their own identifiers (mma-*, *.nlm.nih.gov)

**...but made for machines**

- 5–100 characters long: chars, nums, some punct

- Case-sensitive

- IA’s official identifier, like LCCN, ISBN, ASIN, but more free-form, no structure
Details Page

All items have a Details page

https://archive.org/details/<identifier>

Presents a preview or viewer (if available), relevant metadata, collection membership, uploader, user reviews, & more

Presentation differs depending on type of items (collections, user account, etc.)

Also offers History & Download options
Item directory

Item files stored in directories

https://archive.org/download/<identifier>

Lists all files (and subdirectories) in the item directory (Details only shows interesting ones)

Some files are from uploader, some generated by IA, some derived from other content

Items may move around cluster, but /download/ URL is persistent
Item structure

Item structure is defined by files and metadata

User uploads individual files, tarballs, ZIP, subdirs, etc.

<identifier>_meta.xml: Item metadata, such as title, license, etc.

<identifier>_files.xml: List of all files in the item + size, modification times, checksums, file types, derivation history

Both available via Metadata API
Item metadata

Contains sundries such as identifier, collection, license, title, & more

Applies to item, not any particular file

Metadata elements may be supplied by user, retrieved from content metadata, or generated by IA’s workers or admins

Flexible metadata schema; currently no schema enforcement

Editable @ https://archive.org/editxml/<identifier>
## Formats

Defined by IA for major file formats

“JPEG”, “Animated GIF”, “PDF”, “EPUB”, etc.

Similar in concept to MIME (image/jpeg, video/mp4, etc.)

## Sources

- original
- metadata
- derivative
source=”original”

Original files are content loaded into an item

Uploaded by user (external or an IA account, i.e. Wayback Machine, Archive-It!, etc.)

May be source of other content forms generated by IA (derivatives)

There are exceptions to above (for example, _files.xml and _meta.xml may be originals)
Metadata files pertaining to other content

Item, original file, or derivation metadata

May be generated by IA or fetched from external source

E.g. MARC (library) records, track listings, IA user reviews, etc.

.torrent files are metadata
Derivatives are alternate representations

<original>filename</original> indicates source

Think thumbnails, different encodings (MP3 → OGG), previews (waveforms of audio tracks), PDFs of images, etc.

Created by workers within IA’s cluster

May be deleted and regenerated
Metadata API

Provides item metadata in JSON format

https://archive.org/metadata/<identifier>

Combines _files.xml + _meta.xml + location details

**d1** is primary (PRI) or solo (SOLO), **d2** is secondary (SEC)

**dir** is logical path on all named servers

Permissions

Permissions come from item’s collection & the user

- Item uploader always has item r/w privs
- Users have “privs” (list of collection ids)
- If item’s collection (or any of that collection’s parent collections) appears in user’s privs list, user has r/w
- Unrestricted collections are readable by everyone

Restricted collections control access based on file formats
Permissions: example collections

opensource (& related)

Public unrestricted collections for user uploads
Audio, movies, texts, etc.
If items not added to specific collection, they go to.opensource

printdisabled

<access-restricted> in _meta.xml
<public-format> whitelists content
Items added to this collection get its priv model
Homework

Just archive, baby

Create a unique piece of content—funny image, poem, sound clip

Upload it

Search for your item

Edit your item’s metadata

Share it with your friends
Internet Archive
Content Storage
& Item Structure

THANK YOU!

John Gonzalez • jcg@archive.org • September, 2016
(derived from presentation of Jim Nelson • 11 April 2016)