

Community Media Archive – Advanced Archiving

John Hauser
Special Projects Mgr.
Access Humboldt

2012 ACM National Conference
July 31 3:30pm - 5pm

Vision

- A collection of broadcast quality, locally produced shows with sufficient descriptive metadata from Community Media Centers that can be freely shared.

Scope of Internet Archive

- nonprofit digital library founded in 1996
- 7+ million items (books, videos, audio, live music)
- 150 billion+ web pages archived in the “Wayback Machine”
- Aim is “Universal Access to all Knowledge”

Scope of Community Media Archive

- Almost 4 years old
- 13,000+ videos
- 23 access centers
- 880,000 downloads
- 27TB storage

Related Collections

- Access centers not part of CMA – CMAP, CTN5
- Access centers without their own collection
 - Grand Rapids, Fort Wayne, Maui
- New collections with ACM compliant MPEG2
 - TedTalks - 1,200 videos
 - PopTech - 200 videos

Community Media Archive on the Roku

To add the private Internet Archive channel that contains the Community Media Archive to your Roku box, got to:

<https://owner.roku.com/add/NMJS5>

Getting Started – Conceptual

- Collection, item, file
- “details” page – item level
- Metadata – item level
- Getting organized is more important than technical knowledge

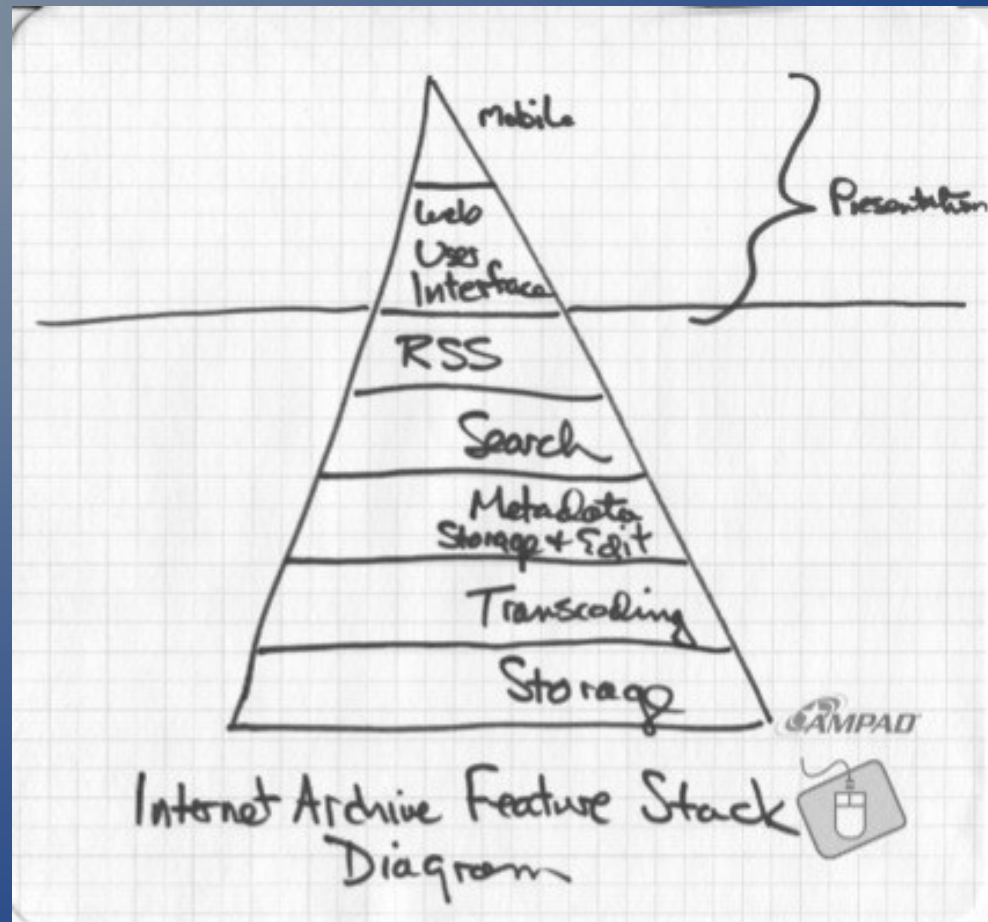
Getting Started - Procedural

- Register an email address
- Email the collections group to add a collection (collections-service@archive.org)
 - Include a descriptive blurb, graphic logo, registered email addresses for admin access
 - Ask that it be setup as a “subcollection” of the `community_media` collection
- input formats they accept and what output formats they derive
- <http://archive.org/help/derivatives.php?mediatype=>

Getting Started – Procedural 2

- Use manual interactive interface to upload first several videos
 - <http://archive.org/create/>
- Bulk uploader available
 - <https://github.com/kngenie/ias3upload>
 - Uses a Comma Separated Value file for metadata

Internet Archive Feature Stack Diagram



“Productionalizing” Archiving

- Moving archiving from an afterthought to part of your production workflow

Production Workflow – SD/SD

- Shoot->Edit->Transcode->Present->Playback Server->Schedule->VOD->Air----maybe sometime in the future someday---->Archive
- Shoot->Edit->Transcode->Present->Playback Server->Schedule->Archive->VOD->Air

Production Workflow – HD/SD

- Shoot->Edit->Present->Archive w MPEG2 profile->Playback Server->Schedule->VOD->Air
- Shoot->Edit->Present->Local Execution of IA's Transcode Process->Playback Server->Schedule->VOD->Air

Metadata and the Archive 1

- Minimal required – identifier, “creator”, title, description
- Anything accepted (and retrievable through `_meta.xml` file on the item's “detail” page
- subject(s) – what shows up under “Browse by subject/keyword” link for collection
- Identifier must be unique across 7+ million items!

Metadata and the Archive 2

- File naming restrictions
 - A-zA-Z0-9._ - no spaces, parenthesis, braces, pound signs, colons allowed in file names
 - Playback server is likely more permissive
 - File suffix matters for animated gif to appear in collection listings

Internet Archive Peculiarities

- What stage is my file in?
 - <http://archive.org/catalog.php?justme=1>
- Every item has a history and it's kept forever!
 - That's good!
 - Runtime and audio/video parameters from ffmpeg output available for every file transcoded
 - http://catalogd.us.archive.org/log_show.php?task_id

Internet Archive Interfaces

- **Advanced Search**
 - Fielded, boolean search (*not* Google style!)
- **Metadata XML and JSON files**
 - [http://www.archive.org/details/\\$identifier&output=json](http://www.archive.org/details/$identifier&output=json)
- **Metadata cleanup after the fact**
 - Single or few items - **Edit Item** interface
 - Many Items – Bulk Uploader in Metadata Merge Mode – no files uploaded, metadata in CSV format
- **RSS feeds**

Internet Archive Admin Interfaces

- “Edit Item”
 - Useful for adding or correcting metadata for a single item
- “Item Manager”
 - “make dark” function
 - “re-derive” certain derivative formats
- “Item History”
 - Links to all prior logs for the item

Finding Programs for Re-Broadcast

- Advanced Search “AND format:MPEG2” search qualifier
- TedTalks, PopTech collections
- Archive.org RSS feeds aren't good enough
 - No runtime reported
 - Subject/keywords not shown
 - Used MRSS 2.0 spec but no one else does

Downloading Programs for Re-Broadcast

- Grab the MPEG2 version!
- Grab the `_meta.xml` file or “screen scrape” title, description, runtime
- Grab the `_files.xml` file; runtime (in seconds) available from the `<length>` element
- Want to be notified when a new episode becomes available?
 - Subscribe to RSS feed for the collection or create a custom one from advanced search interface

Updates to the CMA

- Bulk uploader now **available**
- Access Humboldt upload service!
- New collections available with ACM compliant MPEG2 letterboxed video
 - **Poptech** – 201 videos
 - **TedTalks** – 1,200 videos
- Telvue's “one button” upload to Internet Archive coming as of sw release 3.23

More updates

- Video derivation logic **released** by archive.org
 - Now you can run the same logic they use on your own server
 - Eliminates upload, queue, download delays from your workflow
- HD/SD toggle in archive user interface for web view
 - Letterboxed MPEG2 version for playback still created

Gift Cultures

- <http://www.catb.org/esr/writings/homesteading/hom>
- Reputation is based what you give away
- Imagine Public Access as a Gift Culture
- What will you contribute?