

The Archive Corps

A founding-cohort curriculum for distributed stewardship of the Community Media Archive

Format: 4 live sessions over 6 weeks (via video call), 60–90 min each, plus a short between-session task after sessions 1–3. **Cohort:** 6–8 people, mostly non-technical. **Goal:** by the end, each participant can independently steward a slice of channels and make real, well-described uploads to a live collection — with judgment, not just clicks.

How this curriculum is built

Four design choices shape everything below, so they're worth stating up front.

- **Anchored to the real workflow, not archiving theory.** Principles are taught just-in-time, the moment they matter for a real task — never as abstract front-loaded lecture. Four sessions isn't enough to make an archivist; it is enough to make a confident, judgment-capable contributor to this specific pipeline.
- **The web path, not the command line.** Because the cohort is mostly non-technical and the capability you most need is curation, the whole curriculum uses Internet Archive's browser-based upload and metadata editor. The batch scripts get explained so people know where their work fits — but participants never have to touch a terminal.
- **Practice in a sandbox, graduate to live.** Sessions 1–3 practice in a test collection where mistakes are harmless. Only session 4 touches the live collection, once people have the reps. That's how you get "real" without "risky."
- **The gaps between sessions do the teaching.** Spacing sessions across 6 weeks is deliberate. Each session ends with a small hands-on task; the next opens by working through what people hit. That between-session reps is what converts "I watched a demo" into "I can do this."

The throughline: *a script can download 26,000 videos a month, but only a person can decide whether a video is worth keeping, whether it's described well enough to be found in ten years, and whether it's in the right place. This curriculum trains that human-judgment layer — the part that can't be automated.*

The arc at a glance

Wk	Session	What they can do after	Pillar emphasis
1	Why we save, and what we're part of	Explain the mission and the whole pipeline; navigate the archive; spot good vs. bad metadata in the wild.	Digital archiving
2–3	Describing things so they're findable	Write strong titles, descriptions, dates, subjects; apply a consistent house style; make a first sandbox upload.	Metadata principles
4–5	Inside the Internet Archive	Understand items/collections/identifiers and which fields they can edit; fix broken metadata; upload cleanly to the sandbox.	IA specifics
6	The real thing, supervised	Steward a real slice end-to-end and make supervised live uploads; know how to ask for help and hand off.	All three, applied

Sessions run roughly every other week. The cohort can be a standing email thread between meetings for questions and to share what they're finding.

Purpose

Build shared motivation and a mental map of the whole operation before anyone touches a tool. People stick with volunteer work when they understand why it matters and can see where they fit. This session is the “why” and the “big picture”; the hands-on starts light.

Run of show

1. **Welcome & introductions (10 min).** Each person: name, station, and one community-media item they'd be heartbroken to lose. Sets the emotional stakes.
2. **Why community media is worth saving (10 min).** Provenance, ephemerality, and “who else is keeping this?” Often no one. Public-access and local-government video vanishes when a channel is deleted.
3. **The pipeline, end to end (20 min).** Walk it plainly: 2,100 channels → ~26,000 videos/month downloaded → batch processing → upload to collections. Show the scripts running once, as a demo, so the automated part is demystified — then make clear their role is the human layer, not the scripts.
4. **“Good enough” preservation at volume (15 min).** You can't perfect 26,000 items a month. Introduce the idea of a usable-and-findable baseline vs. archival perfection, and why consistency beats heroics.
5. **Tour of the live archive (15 min).** Browse the Community Media Archive together. Look at a well-described item and a poorly-described one side by side. Ask: which one will someone find in ten years?
6. **Wrap & assign (5 min).**

Concepts introduced

- Provenance and why source/origin matters.
- Ephemerality: why “it's on YouTube” is not preservation.
- The difference between the automated layer (scripts) and the judgment layer (people).

Between-session task: find 3 items already in the Community Media Archive — one described well, one described poorly, one you're unsure about — and jot one sentence each on what makes the metadata work or fail. Bring them to session 2.

Purpose

This is the heart of the cohort — metadata is the craft you most need to distribute, so it gets the longest session and a first real (sandbox) upload. The goal: by the end, everyone has described and uploaded one item in the test collection.

Run of show

1. **Warm-up: share your three items (10 min).** Quick round from the assignment. Surface patterns: what made the good ones good?
2. **The three jobs metadata does (15 min).** Descriptive (what is this?), administrative (where did it come from, who uploaded it, rights), and structural (how parts relate). Keep it concrete — tie each to a field they'll actually fill in.
3. **The fields that matter most (20 min).** Title, description, date, creator, subject/tags, and source. For each: what good looks like, common mistakes, and a house-style rule (e.g., how to format dates, how to title a city-council meeting consistently).
4. **Consistency beats cleverness (10 min).** Why controlled vocabularies and a shared house style make 26,000 items findable. A clever one-off title helps no one; a predictable one helps everyone.
5. **Guided first upload to the sandbox (30 min).** Using a sample video you provide, walk everyone through the browser uploader and metadata editor into the test collection. Everyone does it live, screens shared as needed.
6. **Debrief & assign (5 min).**

Concepts introduced

- Descriptive vs. administrative vs. structural metadata.
- Controlled vocabulary and house style; why consistency is a findability tool.
- The browser upload flow and the metadata editor (first exposure).

Between-session task: upload 2 more sample items to the test collection on your own, applying the house style. Note anything that confused you — those questions open session 3.

Purpose

Now that they can describe and upload, teach how IA actually works underneath — enough to avoid the mistakes that are hard to undo, and to confidently fix existing items. This is where the platform specifics land, grounded in tasks they've already done.

Run of show

1. **Open with their questions (10 min).** Work through the confusions from the between-session uploads. Real friction is the best curriculum.
2. **Items, collections, and identifiers (20 min).** How an item relates to a collection; that the identifier becomes the permanent URL (`archive.org/details/[identifier]`) and can't be casually changed — so naming discipline matters. Why human-readable identifiers beat opaque strings.
3. **What you can edit vs. what's locked (15 min).** Contributors can edit title, description, subjects, and most descriptive fields. Media type, collection, uploader, and identifier are admin-only. Knowing the line prevents wasted effort and panicked emails.
4. **Search and discovery (15 min).** How metadata feeds search — searching `key:value`, why a missing or sloppy field makes an item effectively invisible. This closes the loop on why session 2's discipline matters.
5. **Fixing broken metadata (15 min).** Hands-on: each person finds a weak item in the test collection and improves it. Editing is as valuable as uploading — most stewardship is maintenance.
6. **Wrap & assign (5 min).**

Concepts introduced

- Item / collection / identifier model; permanence of the identifier-URL.
- Editable vs. admin-only fields.
- How metadata drives search; the 15-minute derive/refresh delay so changes don't appear instantly.

Between-session task: pick a real slice you'd want to steward (a region, a set of stations, a topic) and write a few sentences on why. This becomes your starting territory in session 4.

Purpose

Graduation. Move from the sandbox to the live Community Media Archive and have each person make real, supervised uploads to their chosen slice. They leave as working stewards, not trainees.

Run of show

1. **Claim your slice (15 min).** Each person names the territory they wrote up. Resolve overlaps, fill gaps. This is also your living succession roster.
2. **From sandbox to live: what changes (10 min).** The stakes are higher and edits are public. Cover the do-no-harm rules and when to stop and ask rather than guess.
3. **Supervised live uploads (40 min).** Working block: everyone uploads real items to the live collection while you're on the call to catch issues in real time. The safety net is you, present.
4. **How to get unstuck (10 min).** Where to find the house-style guide, the IA help center, and how to ask the cohort email list a good question. Self-sufficiency is the point of stepping back.
5. **What happens next (15 min).** Ongoing cadence, how cohort 1 can help teach cohort 2, and a light check-in rhythm. End on ownership and momentum.

Concepts introduced

- Do-no-harm editing on a live, public archive; when to escalate.
- Self-service support paths and good question-asking.
- The “each cohort trains the next” model that makes the work self-sustaining.

Graduation outcome: each participant owns a named slice and has made real supervised uploads. The cohort becomes the seed of distributed stewardship — and your first teachers for the next group.

What you'll need to prepare

- **A test collection** for safe practice (IA provides test_collection for this purpose, or use a dedicated sandbox you control).
- **A small set of sample videos** — ideally 5–6 with varied, realistic metadata challenges (a clear one, a messy one, an undated one) so practice mirrors reality.
- **A one-page house-style guide** — your conventions for titles, dates, subjects, and source. This is the single most reusable artifact; it outlasts the cohort.
- **Each participant: a free Internet Archive account** created before session 2.
- **Video-call setup with screen sharing**, and ideally a co-host for sessions 2 and 4 so someone can help individuals while you teach.

A note on scope: this trains stewardship and curation — the capability you said matters most. It deliberately does not train people to run or maintain the batch scripts; that's a separate, smaller, more technical track for the rare developer-minded volunteer, and a good candidate for a future cohort once this model is proven.