Generative-Lives-Bluesky [2025] — Project Canvas (v2)

[CSG: Created by ChatGPT-5 in Extended Thinking + Study and Learn Mode, in response to a prompt by Colin Greenstreet, with edits by Colin Greenstreet in square brackets and marked with his initials "CSG"]

[CSG: Purpose:] A living document we'll refine together. Use this as our shared workspace for goals, decisions, questions, models, and next steps.

[CSG: Version: 2.0]

[CSG: Date and time of creation: Tuesday, October 28th 2025, 09:26 AM GMT]

Project overview

- **Purpose:** Explore how to conceptualize, appraise, capture, describe, preserve, and provide access to social media posts (initial focus: Bluesky) from an archivist's perspective.
- **Scope (initial draft):** Public posts and threads; embedded media; interactions (likes, reposts, replies); account-level context.

What we know so far

- **Key decisions to date:**
    - *(placeholder — bring forward any decisions reached in prior threads)*
- **Assumptions:**
    - Social posts can be treated as records; context is critical for interpretation.
    - Threads and account-level production may constitute record series/fonds-like aggregations.

Open questions (for our Socratic exploration)

1. **Unit of description:** Should the primary unit be the individual post, the thread/conversation, or the account's output as a series? What are the appraisal and access trade-offs of each?
2. **Essential metadata:** What is the minimal viable metadata set for usability and long-term preservation? (e.g., identifiers, timestamps, authorship, fixity, relationships, rights)
3. **Context capture:** How much relational context (threads, quotes, embeds, replies) is necessary to preserve evidential value?

4. **Ethics & rights:** What consent, privacy, and community-harm considerations apply to public social media archives?
5. **Authenticity & integrity:** How will we evidence provenance and chain-of-custody across exports, migrations, and normalizations?
6. **Preservation risks:** What failure modes are specific to Bluesky/ATProto (link rot, dereferencing, missing blobs, algorithmic feeds)?

Principles & standards to draw on (working list)

- Records Continuum; Provenance & Original Order; Appraisal theory (documentation strategy, macro-appraisal)
- Description models: **ISAD(G)**, **ISAAR(CPF)**, **RiC-CM** (Records in Contexts), **Dublin Core**
- Preservation: **OAIS** functional model; **PREMIS** events/agents/objects/rights

Candidate data model (first pass)

- **Entities:**
    - **Agent** (Account/Author)
    - **Record** (Post)
    - **Aggregation** (Thread/Conversation; Series for Account output)
    - **Digital Object** (Media/Blob)
    - **Event** (Creation, Edit, Repost, Like, Reply, Capture, Fixity check)
    - **Rights** (Licenses, Terms, Takedowns)
- **Key relationships:** Agent ↔ Record (createdBy); Record ↔ Aggregation (memberOf); Record ↔ Record (repliesTo/quotes/embeds); Record ↔ Digital Object (hasPart); Event ↔ (Record/Digital Object) (actedOn)

Minimal metadata (MVP draft)

- Identifier(s), Stable URI, ATProto record CID & DID
- Author (handle + DID), Timestamp(s) (created/edited), Text, Language
- Structural/context: replyTo, threadRoot, quotedPost, embeds
- Media refs (blob IDs), checksums, file types, sizes
- Capture metadata: capture tool, date, fixity, source URL
- Rights/ethical notes; sensitivity flags; access status

Workflow sketch (alpha)

1. **Appraise & select** (unit = post/thread/account series)
2. **Capture** (export API / atproto; web capture where needed)
3. **Normalize** (JSON ➜ preservation package; map to data model)
4. **Enrich** (link threads; compute checksums; language; entity extraction—optional)
5. **Store** (packages + metadata; versioning)
6. **Access** (catalogue records; viewer; redaction where necessary)
7. **Preserve** (fixity schedule; format migration; re-harvest policies)

Exemplars & precedents (to investigate)

- Social media archiving by national archives, libraries, and NGOs (Twitter/X, Tumblr, Reddit); Web archiving (WARC; Browsertrix; Archive-It)
- Legal/ethical frameworks (jurisdiction-dependent)

Risks & mitigations (draft)

- API/terms volatility → document methods; maintain provenance; prefer exports over scraping where possible
- Context loss → capture thread graphs; preserve quoted/embedded targets
- Link rot/media loss → download blobs where permitted; checksums; replication

Next steps (checklist)

-

To import from previous threads

Place content, decisions, and notes from earlier project conversations here so we can consolidate.