Implement a daemon that corrects out-of-sync cover art and event art metadata on archive.org

Prepared by: Nishant Kumar
Date: 17/3/25

Project Length:175 hours(medium)

Overview

According to ticket IMG-129, I received some data that I need to resolve. After carefully analyzing the issues, I have outlined the solutions below and how we can address them.

---

1. Identify and Categorize Issues

Based on the audit, the key issues are:

- **DeletedItem::** Items that were removed but still exist in the database.
- **MergedItem::** Items that were merged but have leftover files or metadata inconsistencies.
- **EmptyItem::** Releases that exist but have no cover art.
- **Item::** Items that exist but contain metadata mismatches.
- **Metadata::** Mismatched metadata fields in `mb_metadata.xml`.
- **Files::** Missing files, incorrect indexing, or broken image links.
- **CAAIndex::** Corrupt or missing `index.json` files.
- **Item::exists**: Items that are listed but do not exist.

---

2. Resolution Plan for Each Issue Category

1. Reindexing Process (High Priority & Standard)

- **Goal:** Fix metadata inconsistencies, missing `index.json`, and out-of-sync information.
- **Action:** Queue affected releases for reindexing using the CAA-Indexer.
- **Urgency Levels:**

    - **High Priority:** Items with missing or incorrect `index.json` (reindex_high_priority).
    - **Standard:** Metadata mismatches and outdated schemas (reindex).
    - **Extended:** Metadata-only inconsistencies (reindex_w_metadata).

2. Handling Deleted, Merged, and Empty Items

- **Goal:** Clean up redundant or improperly removed data.
- **Actions:**
    - Remove `DeletedItem::*` and `MergedItem::*` items properly (deleted_properly_delete, merged_properly_delete).
    - Investigate and remove `EmptyItem::*` where necessary.

3. File Integrity Checks

- **Goal:** Ensure image files and metadata exist and are properly linked.
- **Actions:**
    - Fix missing or corrupt images (manual_check).

4. Handling Darkened Items

- **Goal:** Mark and handle temporarily disabled content.
- **Actions:**
    - Use darkened_items list to update DB status accordingly.
    - Suggest a long-term solution to automate darkened item tracking.

---

3. Execution Strategy

1. **Batch Processing & Automation**
    - Run reindexing in batches, prioritizing `reindex_high_priority`.
    - Automate API submissions for IA updates.
2. **Manual Interventions Where Necessary**
    - Investigate `manual_check` flagged items.
    - Verify `darkened_items` against IA policies.
3. **Monitoring & Validation**
    - After fixes are applied, re-run audit on sample data.
    - Compare before/after reports to measure success.

---

4. Expected Outcomes

- **Data Accuracy:** Improved metadata integrity across IA and MB.
- **File Consistency:** Resolved missing/corrupt file issues.
- **Searchability:** Better indexing and retrieval efficiency.
- **Reduced Errors:** Eliminated redundant or orphaned entries.