Function 4: Filter CSV for Historical Records (95+ Years Old)

Overview

Filters a CSV export to identify and isolate bibliographic records that are **95 years old or older** based on their date metadata. The cutoff year is calculated dynamically (current year - 95), and dates are rounded down to the year only when applying this age requirement. This is useful for identifying historical/older materials that may require special handling or metadata work in digital collection workflows, particularly for copyright and rights statement purposes.

What This Function Does

1. **Calculates** the cutoff year dynamically: current year - 95 (e.g., in 2026, the cutoff is 1931)
2. **Reads** the most recent `alma_export_*.csv` file (or a specified input CSV)
3. **Scans** all records for date values in these Dublin Core fields:
   - `dc:date`
   - `dcterms:created`
   - `dcterms:issued`
   - `dcterms:dateSubmitted`
   - `dcterms:dateAccepted`
4. **Extracts** any 4-digit year found in those fields (dates rounded down to year only)
5. **Filters** to keep ONLY records where at least one date field contains a year **≤ cutoff year** (95+ years old)
6. **Outputs** a new timestamped CSV file (e.g., `historical_export_20260112_143022.csv`) containing only the records 95+ years old

When to Use This Function

- After exporting records to CSV (Function 3) to identify historical materials
- To find materials that may need special conservation handling
- To identify older works that might have different copyright/rights statements
- To create a subset of materials for targeted metadata enhancement
- For research or analysis of historical collections

How to Use

Single File Processing

1. Run **Function 3: Export Set to DCAP01 CSV** to create an `alma_export_*.csv` file
2. Click **Function 4: Filter CSV for Records 95+ Years Old**
3. The function automatically finds the most recent `alma_export_*.csv` file
4. A new filtered CSV is created with filename: `historical_export_YYYYMMDD_HHMMSS.csv`
   - Example: `historical_export_20260112_143022.csv`

Output File Format

- **Same columns** as the input CSV (all Dublin Core and custom fields preserved)
- **Fewer rows** - only records with at least one date field containing a year ≤ cutoff year (95+ years old)
- All original data is preserved; no fields are modified

Example Results

Example for 2026 (cutoff year = 1931):

Input: `alma_export_20260112_100000.csv` (500 records)

- Record A: dc:date = "1920" → **INCLUDED** (1920 ≤ 1931, 106 years old)
- Record B: dc:date = "1930" → **INCLUDED** (1930 ≤ 1931, 96 years old)
- Record C: dcterms:created = "1925-03-15" → **INCLUDED** (1925 ≤ 1931, 101 years old)
- Record D: dc:date = "1932" → **EXCLUDED** (1932 > 1931, only 94 years old)
- Record E: dc:date = "" (empty) → **EXCLUDED** (no valid date)

Output: `historical_export_20260112_140000.csv` (250 records 95+ years old)

Date Field Processing

The function uses **regular expression matching** to extract years:

- Looks for any 4-digit number matching the pattern `\b(1[0-9]{3}|20[0-9]{2})\b`
- Accepts years from 1000-2099
- **Dates are rounded down to the year only** - only the year component is used for age calculation
- Works with various date formats:
  - `1920`
  - `March 1920`
  - `1920-03-15` (rounded down to 1920)
  - `1920/03/15` (rounded down to 1920)
  - `03-15-1920` (rounded down to 1920)
  - `ca. 1920`

Important: The 95-year age requirement is applied to the extracted year only. A record dated "1931-12-31" is treated as year 1931 for comparison purposes.

Limitations

- **Only filters by year** - The function extracts only the 4-digit year, not the full date (dates rounded down)
- **Dynamic cutoff** - The cutoff year changes each calendar year (e.g., 1931 in 2026, 1932 in 2027)
- **Multiple dates** - If a record has multiple dates, the function checks if ANY are 95+ years old (inclusive OR logic)
- **Text dates** - Dates written as text (e.g., "January Nineteen-Twenty") will NOT be recognized
- **Empty fields** - Records with empty date fields are automatically excluded
- **File location** - The function looks for `alma_export_*.csv` files in the current working directory

Related Functions

- **Function 3**: Export Set to DCAP01 CSV (creates the input CSV)
- **Function 8**: Export dc:identifier CSV (for identifier-based filtering)
- **Function 9**: Validate Handle URLs (for checking specific record links)

Technical Notes

- The CSV reader is **case-insensitive** when looking for date column headers
- All original CSV columns are preserved in the output
- The function is **non-destructive** - original CSV files are never modified
- UTF-8 encoding is used for both input and output files
- **Cutoff year is calculated dynamically** each time the function runs
- **Year-based comparison**: Dates are rounded down to year only (month/day ignored)

The 95-Year Rule

Why 95 Years?

The 95-year threshold is used in copyright and rights management:

- Aligns with U.S. copyright law for works published 1923-1977 without proper notice
- Materials 95+ years old typically fall into the public domain
- Provides a clear, objective criterion for identifying historical materials eligible for public domain designation
- Conservative approach that safely identifies materials past copyright protection

Dynamic Calculation

The cutoff year is calculated as: **current_year - 95**

Examples:

- **2026**: cutoff = 1931 (includes 1931 and earlier)
- **2027**: cutoff = 1932 (includes 1932 and earlier)
- **2030**: cutoff = 1935 (includes 1935 and earlier)

This means:

- The function automatically adjusts each year
- No code changes needed as time passes
- Consistent application of the 95-year rule

Date Rounding

Important: Dates are rounded **down** to the year only:

- "1931-01-01" → year 1931 → **INCLUDED** in 2026 (exactly 95 years)
- "1931-12-31" → year 1931 → **INCLUDED** in 2026 (95 years old)
- "1932-01-01" → year 1932 → **EXCLUDED** in 2026 (only 94 years old)

This conservative approach ensures all materials from a given year are treated consistently, regardless of the specific month/day.

Recent Updates

January 2026 Enhancement

Changed from hard-coded 1930 to dynamic 95-year calculation

What Changed:

- Previously filtered for dates before 1930 (hard-coded)
- Now filters for records 95+ years old (dynamic calculation)
- Cutoff year updates automatically each calendar year
- Dates rounded down to year only for age comparison

Why This Matters:

- Aligns with U.S. copyright law (95 years for works published without proper notice)
- No manual updates needed as years progress
- More meaningful criterion (age-based) than arbitrary year
- Consistent with rights statement workflows and legal requirements

Impact:

- In 2026: now filters for ≤1931 (previously ≤1929)
- Output filenames are now year-agnostic: `historical_export_...`
- Precisely targets materials that have entered the public domain
- Filename remains consistent as years progress (no year in filename)