Function 4: Filter CSV for Pre-1931 Dates

Overview

Filters a CSV export to identify and isolate bibliographic records with dates **BEFORE 1931**. This is useful for identifying historical/older materials that may require special handling or metadata work in digital collection workflows.

What This Function Does

1. **Reads** the most recent `alma_export_*.csv` file (or a specified input CSV)
2. **Scans** all records for date values in these Dublin Core fields:
   - `dc:date`
   - `dcterms:created`
   - `dcterms:issued`
   - `dcterms:dateSubmitted`
   - `dcterms:dateAccepted`
3. **Extracts** any 4-digit year found in those fields
4. **Filters** to keep ONLY records where at least one date field contains a year **< 1931**
5. **Outputs** a new timestamped CSV file (e.g., `pre1931_export_20240115_143022.csv`) containing only the pre-1931 records

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 Pre-1931 Dates**
3. The function automatically finds the most recent `alma_export_*.csv` file
4. A new filtered CSV is created with filename: `pre1931_export_YYYYMMDD_HHMMSS.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 < 1931
- All original data is preserved; no fields are modified

Example Results

Input: `alma_export_20240115_100000.csv` (500 records)

- Record A: dc:date = "1920" → **INCLUDED** (1920 < 1931)
- Record B: dc:date = "1950" → **EXCLUDED** (1950 ≥ 1931)
- Record C: dcterms:created = "1925-03-15" → **INCLUDED** (1925 < 1931)
- Record D: dc:date = "" (empty) → **EXCLUDED** (no valid date)

Output: `pre1931_export_20240115_140000.csv` (250 records with pre-1931 dates)

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
- Works with various date formats:
  - `1920`
  - `March 1920`
  - `1920-03-15`
  - `1920/03/15`
  - `03-15-1920`
  - `ca. 1920`

Limitations

- **Only filters by year** - The function extracts only the 4-digit year, not the full date
- **Multiple dates** - If a record has multiple dates, the function checks if ANY are pre-1931 (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