Wrangling and Refining "Kultura" Metadata: Complete Step-by-Step Guide

Project Setup

1. Import and Create Project

Import your spreadsheet into OpenRefine

Create a project using descriptive naming:

Format: kultura_[start_year]_[end_year]

Example: kultura_1948_1950

Data Cleaning: Converting Scraping Residue

2. Transform Web Scraper Columns

2.1 Journal Column (from web-scraper-order)

Rename column: Dropdown → Edit column → Rename column → journal

Clear existing values: Dropdown → Edit cells → Common transforms → To null

Add journal name: Dropdown → Facet → Text facet → Edit → Enter Kultura

2.2 Publisher Column (from web-scraper-start-url)

Rename column: Dropdown → Edit column → Rename column → publisher

Clear existing values: Dropdown → Edit cells → Common transforms → To null

Add publisher name: Dropdown → Facet → Text facet → Edit → Enter Instytut Literacki

Adding New Columns

3. Add Essential Metadata

3.1 Realm Column

Create column: Dropdown → Edit column → Add column based on this column

Rename: Dropdown → Edit column → Rename column → realm

Move to beginning: Edit column → Move column to beginning

Set value: Facet → Text facet → Edit → Enter PL

3.2 Volunteer Column

Select the last column: Dropdown → Add column based on this column

Rename: Dropdown → Edit column → Rename column → volunteer

Clear values: Dropdown → Edit cells → Common transforms → To null

Add your name: Select first row → Edit → Enter your name → Apply to all identical cells

Table of Contents Processing

4. Clean and Structure the Table of Contents

4.1 Initial Cleanup

Rename column: Dropdown → Edit column → Rename column → table_of_contents

Remove whitespace:

Dropdown → Edit cells → Common transforms → Collapse consecutive whitespaces

Dropdown → Edit cells → Common transforms → Trim leading and trailing whitespaces

4.2 Extract Issue Information

The table_of_contents is considered multivalued since it contains both issue info and content info.

⚠️ *Analyze the data to recognize consistent patterns, then apply these patterns as rules for column splitting, cell editing, and data manipulation.*

Split by issue marker:

Dropdown → Edit column → Split into several columns

Separator: 1 SPIS RZECZY

This separates issue info from content

4.3 Create an issue_id column:

Rename the new column to issue_id

Clean residue using the replace function: Dropdown → Edit cells → Replace (remove unwanted text) or manually.

4.4 Create Year and Issue Columns

Add column based on issue_id: Dropdown → Edit column → Add column based on this column

Split by date separator: Dropdown → Edit column → Split into several columns

separator: /

Rename columns:

First column: Dropdown → Edit column → Rename column → year

Second column: Dropdown → Edit column → Rename column → issue

4.5 Format Issue Column

Clean whitespace: Dropdown → Common transforms → Remove leading and trailing whitespace

Standardize format: Ensure single digits have a leading zero (9 → 09)

Convert to text: Dropdown → Edit cells → Common transforms → To text

Proofread for consistency

4.6 Finalize issue_id Format

Replace / with nothing: Dropdown → Edit cells → Replace / with empty field

Replace - with _: Dropdown → Edit cells → Replace - _

Final format examples:

Single issue: 197506 (year + issue)

Double issue: 197506_07 (year + issue_issue)

Content Extraction and Author Processing

5. Split Content Entries

5.1 Split Multivalued Cells

Pattern Recognition: Each entry starts with a page number

Split using regex: Dropdown → Edit cells → Split multivalued cells

Separator: \s+\d+\s+

✅ Check "Regular expression"

Regex Explanation:

\s+ = one or more whitespace characters

\d+ = one or more digits

\s+ = one or more whitespace characters

5.2 Remove Section Headers/Rubrics

Identify rubric rows (section headers like "KSIĄŻKI", "VARIA")

Delete rubrics: Select rubric row → Edit → Delete → Apply to all identical cells

⚠️ *Important: Proofread carefully — rubrics vary and don't always repeat exactly*

6. Extract Author Information

6.1 Create Author Column

Based on table_of_contents: Dropdown → Edit column → Add column based on this column

Split by author-title separator:

Dropdown → Edit column → Split into several columns

Separator: : (colon separates author from title)

6.2 Clean Author Data

Remove title column: Dropdown → Edit column → Remove column (keep only author column)

Rename: Dropdown → Edit column → Rename column → published_authors

Standardize formatting:

Dropdown → Edit cells → Common transforms → To Titlecase

Remove page numbers: Dropdown → Edit cells → Replace (or manually)

Use the Text facet to identify and fix inconsistencies:

Dropdown → Facet → Text facet

6.3 Consolidate Authors

Join multivalued cells: Dropdown → Edit cells → Join multivalued cells

Separator: ; (semicolon + space)

Quality Control and Export

7. Final Cleanup and Export

7.1 Remove Empty Rows

Filter non-empty rows:

Select All column → Facet → Facet by blank

Select false to show only non-empty rows

7.2 Export Data

Export → Excel (or preferred format)

Only filtered (non-empty) rows will be exported

⚠️ Data Quality Checklist

- Manually review and edit entries/remove non-content entries
- Confirm all content entries are preserved (no data loss during splitting)

Make sure that:

- All column names are descriptive and consistent
- Date formats are standardized
- Author names use consistent capitalization and overall consistency
- Issue IDs follow the pattern
- No empty rows in final export
- All multivalued cells are properly split and rejoined