

# 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]\_clean

Example: kultura\_1948\_1950\_clean

## 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 Editor-in-Chief Column

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

Rename: Dropdown → Edit column → Rename column → editor\_in\_chief

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

Add your name: Select first row → Edit → Enter Jerzy Giedroyc → Apply to all

identical cells

#### 3.3 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 column 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.*

We can see how the content and issue info are separated by “SPIS RZECZY”. So, we can use it for column splitting.

The pattern here is space, digit, space, SPIS RZECZY, space

Split by issue marker with regular expression:

Dropdown → Edit column → Split into several columns

Separator: \s+\d+\s+SPIS RZECZY\s+

✅ Check "Regular expression"

Regex breakdown:

\s+ = one or more whitespace characters

\d+ = one or more digits

\s+ = one or more whitespace characters

SPIS RZECZY = first line in the actual table of contents

\s+ one or more whitespace characters

Rename the new column to issue\_info: Dropdown → Edit column → Rename column → issue\_info

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

#### 4.3 Create Year and Issue Columns

Split this column by regex separator: Dropdown → Edit column → Split into several columns

Separator: /(=?\d{2}\d{3})

ex. 1991/01/520 - 02/521

☒ Check "Regular expression"

Regex breakdown:

Pattern: year/issue number within a year/issue number over the journal history

\d{2} - exactly two digits

/ - a forward slash

\d{3} - exactly three digits

Rename columns:

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

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

Remove extra whitespaces in the issue column with regex:

Dropdown → Edit cells → Replace

Replace \s\*-\s\* with –

☒ Check "Regular expression"

Regex breakdown:

\s\* whitespace

- hyphen

\s\* whitespace

#### **4.4 Create issue\_id column:**

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

Name it issue\_id

Merge 2 columns:

Select the issue\_id column: Dropdown → Join columns → Select issue column →

Put \_ as separator

Replace – with \_ in issue\_id column: Dropdown → Edit cells → Replace

ex: 1991, 01/520-02/521 → 1991\_01/520\_02/521

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

Proofread for consistency

## **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 Adding rubric column

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

Rename: Dropdown → Edit column → Rename column → rubric

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

Copy rubric name and paste it in the rubric column next to the first entry of this rubric

While working with rubric columns, proofread the table\_of\_contents column. There might be some residue from the splitting; double-check with the journal's table of contents.

## 6. Extract Author and Title Information

### 6.1 Create Author and Titles Columns

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)

Enter 2 in the split into field

Split into \_ columns at most (leave blank for no limit)

### 6.2 Clean Author Data and Titles Data

Rename: Dropdown → Edit column → Rename column → published\_authors

Rename: Dropdown → Edit column → Rename column → published\_works

Standardize formatting:

Dropdown → Edit cells → Common transforms → To Titlecase

Use the Text facet to identify and fix inconsistencies:

Dropdown → Facet → Text facet

## Quality Control and Export

### 7. Final Cleanup and Export

#### 7.1 Exporting Complete Bibliography Index

##### Remove Empty Rows

Filter non-empty rows:

Select All column → Facet → Facet by blank

Select false to show only non-empty rows

Select issue\_id column, use facet

Dropdown → Facet → Text facet

To fill down rubrics for a specific issue:

Dropdown → Edit cells → Fill down

Repeat for each issue\_id

Select issue\_id

For columns publisher, issue\_id, year, issue, cover:

Use the fill down function to fill out the empty cells of the row

Dropdown → Edit cells → Fill down

Export → Custom Tabular → Deselect realm, table\_of\_contents, volunteer fields →

Download tab → xlsx format → Facet → Download

#### 7.2 Exporting Metadata for the Website

Select Undo and cancel your last edit (fill down)

Merge published\_authors rows and published\_titles rows:

Dropdown → Edit cells → Join multivalued cell

Separator ; (semicolon + whitespace)

##### Remove Empty Rows

Filter non-empty rows:

Select All column → Facet → Facet by blank

Select false to show only non-empty rows

##### Export Data

Export → Custom tabular → Deselect All → Deselect title and author fields →

Download tab → xlsx format → Download

Only filtered (non-empty) rows and selected 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