How to Create Tables of Historical Administrative Units

Objectives

Our immediate goal is to collect data on administrative-territorial changes in Soviet Ukraine. The downstream analytical goal is to better understand why countries redraw their internal administrative borders, and what sorts of political, economic and social legacies these changes leave behind.

There are several types of boundary changes: create, merge, split, abolish. These changes can apply to legislative, jurisdictional, and administrative borders. These changes happen for a variety of reasons, from technocratic "optimization" and demographic changes, to political survival.

Like many countries, the Soviet Union frequently changed its internal administrative boundaries throughout its existence, driven by political, economic, and ethnic considerations. These boundary changes varied in the extent to which pre-existing communities were kept intact between the old and new maps. For example, the USSR sometimes consolidated pre-existing political communities into larger units (Chechen-Ingush ASSR), but other times carved them up between neighboring provinces, wiping away all internal borders, leaving no trace of their existence (Volga German ASSR).

We will assemble data on these changes using declassified Soviet gazetteers. A gazetteer is a geographical dictionary or directory that provides detailed information about places, including names, locations, administrative divisions, and sometimes historical or cultural details. Ideally, we will be able to cover the full period of Soviet Ukrainian history from the 1920s to 1991. Our first priority will be to collect data on the pre-WWII period, 1921-1939.

Below is a set of instructions on how to create tables of historical administrative units from scanned PDFs of declassified archival gazetteers.

Step 1: Pick a Gazetteer

1. **Familiarize Yourself with the File and Folder Structure**:
    - We will store the original PDFs in the `YZRA/Data/ATD/Raw` directory in Dropbox
    - Ensure you have access to the scanned PDFs of declassified gazetteers for Soviet Ukraine (called "administrative-territorial division", or *адміністративно-територіальний поділ* in Ukrainian and *административно-территориальное деление* in Russian).
    - These files have names like `YEAR_FileDescription.pdf`, so that sorting them alphabetically also sorts them chronologically
2. **Pick a gazetteer to digitize**
    - Use our team's Trello board to claim a PDF file, and drag its card to the "Working" column

Step 2: Set Up a Spreadsheet for Each Gazetteer

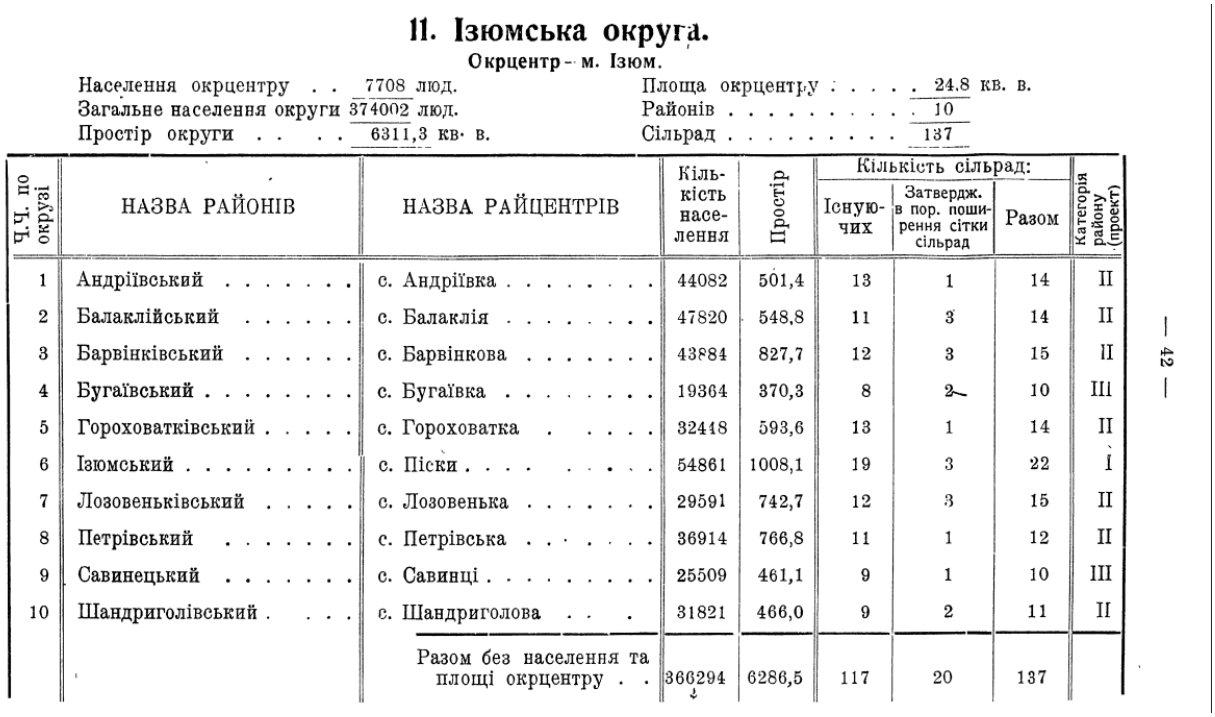
1. **Create a Spreadsheet**:
    - There are 2 types of spreadsheet templates in the `YZRA/Data/ATP/Templates` folder:
        1) `small`: tables of admin-2 units (e.g. districts, counties, rayons)
        2) `big`: tables of admin-3 units (e.g. town, villages)
    - The `small` files are the main priority at the moment. The `big` files are harder to make. Not all gazetteers have information down to the admin-3 level, and even if they do, it's easier to build this kind of table after first creating the `small` one.
    - Open a spreadsheet tool such as Microsoft Excel, LibreOffice Calc, or Apple Numbers.

- Create columns:
  - `year` (year of creation/reorganization),
  - `name_1` (admin-1 unit),
  - `center_1` (capital of admin-1),
  - `name_2` (admin-2 unit),
  - `center_2` (capital of admin-2),
  - `name_3` (admin-3 unit, **"big" tables only**),
  - `name_1_previous` (if applicable),
  - `name_2_previous` (if applicable).
- Include additional columns if relevant information is available in gazetteer (e.g., population, distance to railroad, area).
- **File Naming Rule**: Each spreadsheet must match the name of its source PDF. For example:
  - Source PDF: `1925_Gazetteer.pdf`
  - Spreadsheet: `1925_Gazetteer_small.csv`

2. **Preview the "Final Product"**
   - For an example of what a completed table for a gazetteer might look like, check out the file `1935_AdmTerSSSR_RSFSR_small.csv` in the `Templates` folder. This is a file for Soviet Russia, not Ukraine, but theis is consistent with what a `small` file would look like.
   - This was created with the help of OCR, the code and work files for which are in the `Working` folder.

Step 3: Inspect the Gazetteer

1. **Understand the Structure**:
   - Skim through your PDF to understand its structure, language, and formatting.
   - Identify key sections listing administrative units, their names, centers, and any reference to the year of creation or reorganization for each administrative unit.
   - Most of the information you need will be in tabular or list form, enumerating which rayon belongs to which oblast' (or okruh, or guberniya, depending on the year), like Figure 1.
2. **Assess OCR Feasibility**:
   - Determine whether OCR is feasible based on print quality. For clean, high-quality scans, OCR can save time. Poor-quality scans will require manual transcription.
   - If you are **able to select the text on the PDF page**, this means OCR has already been implemented, and you can save time by copying and pasting the text, instead on manually typing the information in.
   - If you are **not able to select the text**, this means OCR has not yet been (successfully) implemented. If this is the case, you can try running OCR software yourself. But some documents have such bad image quality that even this may not be an option.
3. **Choose OCR Software** (if necessary/feasible):
   - Available tools include:
     - **Tesseract** (free and open-source).
     - **Adobe Acrobat Pro** (paid, user-friendly).
     - **ABBYY FineReader** (specialized for historical documents).
     - **Google Drive OCR** (basic but effective for clean text).
   - For Cyrillic text, ensure the OCR software supports Russian/Ukrainian scripts.
4. **Run OCR**:
   - Process each PDF using OCR software to extract text.
   - Save the output to the `YZRA/Data/ATD/Working` directory, as plain text or structured formats like CSV/Excel for easier editing.
5. **Verify OCR Accuracy**:
   - Manually review OCR results to correct errors, especially in names, administrative terms, and dates.



11. Ізюмська округа.

Окрцентр – м. Ізюм.

Населення окрцентру . . 7708 люд. Площа окрцентру . . . . 24,8 кв. в.
Загальне населення округи 374002 люд. Районів 10
Простір округи . . . . 6311,3 кв. в. Сільрад 137

№ по ч. ч. округі	НАЗВА РАЙОНІВ	НАЗВА РАЙЦЕНТРІВ	Кількість населення	Простір	Кількість сільрад: Існуючих	Затвердж. в пор. поширення сітки сільрад	Разом	Категорія району (проект)
1	Андріївський	с. Андріївка	44082	501,4	13	1	14	II
2	Балаклійський	с. Балаклія	47820	548,8	11	3	14	II
3	Барвінківський	с. Барвінкова	43884	827,7	12	3	15	II
4	Бугаївський	с. Бугаївка	19364	370,3	8	2	10	III
5	Гороховатківський	с. Гороховатка	32448	593,6	13	1	14	II
6	Ізюмський	с. Піски	54861	1008,1	19	3	22	I
7	Лозовеньківський	с. Лозовенька	29591	742,7	12	3	15	II
8	Петрівський	с. Петрівська	36914	766,8	11	1	12	II
9	Савинецький	с. Савинці	25509	461,1	9	1	10	III
10	Шандриголівський	с. Шандриголова	31821	466,0	9	2	11	II
		Разом без населення та площі окрцентру	366294	6286,5	117	20	137	

— 42 —

Figure 1: Page from 1925 gazetteer (`1925_adminterpodil.pdf`)

Step 4: Input Data into Spreadsheets

1. **Read and Extract Data**:
   - Carefully read through the gazetteer sections listing administrative units (admin-2 rayons in the case of `small` tables, and admin-3 populated places in the case of `big` tables).
   - Input extracted data into your spreadsheet row by row for each administrative unit.
   - Include additional columns if relevant information is available (e.g., population, area).
2. **Keep Original Language and Spelling**:
   - Preserve the original language (Ukrainian and/or Russian) and spelling.
   - Do not translate or transliterate names into other languages or scripts.
   - This ensures historical accuracy and consistency with archival sources.
3. **Capture Year Information**:
   - If the gazetteer indicates the year when an administrative unit was created or reorganized, record this in a separate column (`year`).
   - If no year is provided, leave the field blank or mark it as “unknown.”
4. **Track Name Changes**:
   - Many gazetteers include sections at the back enumerating name changes for administrative units.
   - If they have this information, use the columns `name_1_previous` and `name_2_previous` to record historical names where applicable, documenting transitions over time.
5. **Ensure Consistent Spelling Across Years**:
   - Try to standardize spelling of administrative unit names (`name_1`, `name_2`) across all years while keeping original language intact.
   - Pay attention to regional variations in spelling between Ukrainian and Russian (e.g., “Київ” vs. “Киев”).
   - If inconsistencies arise, create a standardized name list or key for reference.
6. **Save Files Regularly and Properly**:
   - Save working spreadsheets with file names matching their source PDFs (e.g., `1925_Gazetteer_small.xlsx`

while working).
- When finalized, export them as `.csv` files with UTF-8 encoding using the same file name (e.g., `1925_Gazetteer_small.csv`).

Step 9: Document Metadata

1. **File Format**:
   - Use plain-text files (`.txt`) for metadata to ensure simplicity and long-term accessibility.
   - Save files in UTF-8 encoding to preserve Cyrillic characters.
2. **File Naming Conventions**:
   - Use consistent names matching those of source PDFs and spreadsheets. Example:
     - `Ukraine_1925_Gazetteer_Metadata.txt`
3. **Minimal Example of Metadata File**
   - `Ukraine_1925_Gazetteer_small_Metadata.txt`:

```
Source: Ukraine_1925_Gazetteer.pdf
Region: Ukraine
Year: 1925
Encoding: UTF-8
Notes: Extracted from declassified gazetteer;
       includes admin-1 and admin-2 units with original spelling.
```

4. **Folder Organization**:
   - Store metadata files in the same folder as their corresponding data tables (`Processed`).

Step 10: Archive and Share

1. **Working Files**:
   - Store spreadsheets currently being developed in the `YZRA/Data/ATD/Working` directory in Dropbox.
   - Ensure file names match their source PDFs.
2. **Completed Files**:
   - Once a file is finalized, save it as a `.csv` file with UTF-8 encoding in the `YZRA/Data/ATD/Processed` directory in Dropbox using a matching file name.
3. **Local Backups**:
   - You are encouraged to make local backup copies of both working and completed files as needed.
4. **Folder Organization**:
   - Maintain a clear hierarchy in Dropbox:
     - `YZRA/Data/ATD/Raw`: For original PDFs of gazetteers.
     - `YZRA/Data/ATD/Templates`: For templates of `small` and `big` tables.
     - `YZRA/Data/ATD/Working`: For in-progress spreadsheet files matching source PDF names.
     - `YZRA/Data/ATD/Processed`: For finalized `.csv` files ready for analysis or sharing.

By ensuring that spreadsheet file names always match their source PDFs across all stages (Raw, Working, and Processed) alongside clear naming conventions, you maintain traceability while preserving consistency throughout your workflow.

Additional Tips

Tip 1: Keep Everything Updated

1. Keep track of your and the team's progress with the spreadsheet `atp_tracker.csv` in the `YZRA/Data/ATP` folder. When you begin working on a file (and have created a working table),

change the value in the `working` column for that file from `N` to `Y`. Save and close. Then do the same for the `processing` column when you're done.
2. Also keep things up to date on our Trello board. Move the card for each file from `To Do` to `Doing` and `Done` as you go.

Tip 2: Preserve Cyrillic Characters

1. **Ensure Proper Encoding**:
   - Always save files with UTF-8 encoding to preserve Cyrillic characters accurately.
   - In Excel, LibreOffice Calc, or Numbers, explicitly select UTF-8 encoding when exporting to CSV format.
2. **File Format Recommendations**:
   - Use `.csv` as the preferred format for final storage because it is widely supported and lightweight.
   - While editing or working on files, `.xlsx`, `.ods`, or `.numbers` formats are acceptable as long as UTF-8 encoding is preserved during export.