How to Digitize Tables of Historical Administrative Units with Large Language Models on Perplexity AI

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, using generative AI.

Step 0. Set Up a Perplexity Pro Account

1. **Visit the Website**: Navigate to Perplexity.ai.
2. **Click "Sign Up"**: Look for the blue "Sign Up" button on the homepage.
3. **Choose Sign-Up Method**:
    - Use your university email (e.g., `netid@georgetown.edu`) by selecting "Continue with Email."
    - NOTE: While you can also sign up via Google or Apple accounts, students with verified `.edu` addresses are automatically upgraded to the "Perplexity Pro" tier for one month or longer when they sign up, and are eligible for a discounted rate of $4.99/month when the trial period expires.
4. **Verify Email**: Check your inbox for a verification email and follow the link to confirm your account.
5. **Log In**: Once verified, log in to Perplexity.ai.

Step 1: Pick a Gazetteer

1. **Familiarize Yourself with the File and Folder Structure**:
    - 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).
    - The PDFs of the original gazetteers are in the `YZRA/Data/ATD/Raw/AI_Ready` directory in Dropbox

    - These files have names like `YEAR_FileDescription_01.pdf`, so that sorting them alphabetically also sorts them chronologically
        - Each gazetteer is split into multiple parts (`_01`, `_02`, etc.), due to memory and response length limits on Perplexity
        - You will need to digitize all parts of each gazetteer
2. **Pick a gazetteer to digitize**
    - Use our team's Trello board to claim a gazetteer, and drag its card to the "Working" column

Step 2. Enter Query and Upload File

1. **Open Perplexity**
    - Log in to your Perplexity account.
    - On the left-hand menu, click **Start New Thread**.
2. **Copy the Query Text for Your Gazetteer**:
    - Due to differences in page layout, each gazetteer requires a slightly different query
    - Select and copy the query corresponding to the gazetteer you are working on:
    - `1921_SpVolUSSR_*`

```
Please convert the attached PDF list of governorates, districts and settlements into
a csv table. Column names should be in English snake_case. Table contents should
preserve original Cyrillic characters.
```

    - `1925_adminterpodil_*`

```
Please extract the contents of the table(s) of regions and districts from the attached
PDF into csv table(s). Column names should be in English snake_case. Table contents
should preserve original Cyrillic characters.
```

    - `1926_TerAdmPodSSSR_*`

```
Please convert the attached PDF list of districts and district centers (grouped by
region) into a csv table. Column names should be in English snake_case. Table contents
should preserve original Cyrillic characters.
```

    - `1930_AdmTerDelSSSR_*`

```
Please extract the contents of the table(s) of regions and districts from the
attached PDF into csv table(s). Column names should be in English snake_case. Table
contents should preserve original Cyrillic characters.
```

    - `1931_AdmTerDelSSSR_*`

```
Please extract the contents of the table(s) of regions and districts from the
attached PDF into csv table(s). Column names should be in English snake_case. Table
contents should preserve original Cyrillic characters.
```

    - `1933_AdmTerPod_*`

```
Please convert the attached PDF list of district characteristics and regions into a
csv table. Column names should be in English snake_case. Table contents should
preserve original Cyrillic characters.
```

    - `1935_AdmTerSSSR_*`

```
Please extract the contents of the table(s) of regions and districts from the
attached PDF into csv table(s). Column names should be in English snake_case. Table
contents should preserve original Cyrillic characters.
```

    - `1936_RayUSRR_*`

```
Please convert the attached PDF table of contents into a csv table of regions and the
districts/okrugs they contain. Column names should be in English snake_case. Table
contents should preserve original Cyrillic characters.
```

    - `1937_AdmTerDelSSSR_*`

```
Please extract the contents of the table(s) of districts and towns (grouped by
region) from the attached PDF into csv table(s). Column names should be in English
snake_case. Table contents should preserve original Cyrillic characters.
```

    - `1940_AdmTerDelSSSR_*`

```
Please extract the contents of the table(s) of districts and towns (grouped by
```

`region) from the attached PDF into csv table(s). Column names should be in English snake_case. Table contents should preserve original Cyrillic characters.`

- Copy the selected text into the search box, but **do not press submit yet**.

3. **Upload the PDF**
   - Press the "Attach image, text or PDFs" button in the lower-right corner of the search box (the icon should look like a paperclip or piece of paper).
   - Start by navigating to the first PDF from your set (e.g. `1936_RayUSRR_01.pdf`).
   - Click "Open".
4. **Set Model and Sources**:
   - Ensure **Auto Mode** is selected (default setting).
   - Click on the "Filters" icon below the search bar.
   - Select **Web and Academic Sources Only**.
5. **Submit the Query**
   - Make sure that both the text and the file have been input into the search box.
   - Press "Enter" or click the search icon (circle with right arrow) to start the thread.

Step 3. Export Answer as a Markdown File

1. **Check the Response**:
   - After you submit the query, a new page will open, with your query followed by a the model's response.
   - Take a look at the response to make sure it actually contains the requested table. The table content should look something like this in the response box:

```
text
region,district_okrug
КИЇВСЬКА ОБЛАСТЬ,Київська м/р
КИЇВСЬКА ОБЛАСТЬ,Житомірська м/р
КИЇВСЬКА ОБЛАСТЬ,Андрушівський
КИЇВСЬКА ОБЛАСТЬ,Бабанський
КИЇВСЬКА ОБЛАСТЬ,Базарський
КИЇВСЬКА ОБЛАСТЬ,Баришівський
КИЇВСЬКА ОБЛАСТЬ,Березанський
КИЇВСЬКА ОБЛАСТЬ,Білоцерківський
...
```

   - This is what a CSV file looks like as plain text: column headings and cell values separated by commas. Actual content will depend on what's in the PDF.
   - If there is no such output in the response, try regenerating the answer with a different LLM:
     - Locate and click on the "Rewrite" button at the bottom of the response box.
     - Select a different LLM (e.g. `Sonar`, `Claude`, `Gemini`) and wait for Perplexity to re-run the query.
     - Note that not all LLM are equally capable of executing this task. For example, instead of getting a table, you may get instructions on how you can make one yourself, with code snippets. If you get such a result, choose a different model and run the query again.
   - If the response contains the requested table, and everything looks about right (table not truncated, etc.), proceed to next step.
2. **Locate Export Option**:
   - Scroll to the bottom of the response box.
   - Find and click on the "Export Answer" button (usually labeled with an export icon).
3. **Save Markdown File**:
   - Choose Markdown format. This will preserve both the text of the query, and the full response, including embedded tables.
   - Save the text file to the directory `YZRA/APT/Data/Processed`
   - **File Naming Rule**: The markdown file must *match the name of its source PDF*. For example:
     - Source PDF: `1936_RayUSRR_01.pdf`
     - Markdown document: `1936_RayUSRR_01.md`

Step 4. Repeat Steps 2-3 with the Next PDF from the Same Set

1. **Start a New Thread**, as described in Step 2.1.
    - Each PDF should be submitted as a new thread, rather than as a follow-up question within the previous thread.
2. In Step 2.3, **Upload the Next PDF** for the gazetteer
    - For example, `1936_RayUSRR_02.pdf`
    - The query text should be the same as what you used for the first file from the set, per Step 2.2.
3. **Export the Next Response as Markdown**
    - Run Steps 2.4-3.3 as before
    - The name for the next Markdown file should again match the name of its source PDF:
        - Source PDF: `1936_RayUSRR_02.pdf`
        - Markdown document: `1936_RayUSRR_02.md`
4. **Repeat** these steps again for all remaining files from your gazetteer (e.g. `1936_RayUSRR_03.pdf`, `1936_RayUSRR_04.pdf`, etc.)

Step 5: Archive and Share

1. **Exported Markdown Files**:
    - All Markdown exports should be saved to the `YZRA/Data/ATP/Processed` directory in Dropbox using a matching file name.
    - Store any intermediate working documents (if you have any) in the `YZRA/Data/ATP/Working` directory.
2. **Local Backups**:
    - You are encouraged to make local backup copies of both working and completed files as needed.
3. **Folder Organization**:
    - Maintain a clear hierarchy in Dropbox:
        - `YZRA/Data/ATD/Raw/AI_Ready`: For original PDFs of gazetteers.
        - `YZRA/Data/ATD/Working`: For in-progress files (in case you have any).
        - `YZRA/Data/ATD/Processed`: For finalized `.md` 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. Preview the "Final Product"

1. For an example of what a properly-generated Markdown export might look like, check out the files `1936_RayUSRR_01.md` and`1936_RayUSRR_02.md` in the `Processed` folder.

Tip 2: 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.