README for the Repository:
GCAP Public Security-Level Data on U.S. Fund Holdings

The Global Capital Allocation Project

November 2025

1. Introduction

This README documents the code and outputs of the N-PORT dataset processing repository. The project constructs a research-ready security–fund–quarter dataset from the U.S. SEC Form N-PORT public flat files by downloading, processing, cleaning, and exporting quarterly Stata DTA files. See Cavani et al. (2025) for additional description, benchmarking, and an application.

2. Structure of Processing Code

The processing pipeline follows a sequential structure with both Python and Stata do-files. To understand a specific step, start with `NPORT_Master.do`, which implements the various processing steps, and trace to the relevant script.

Pipeline Stages

1. **Data Download** (`A_download_nport.sh`): Downloads raw N-PORT filings from SEC sources.
2. **Initial Processing** (`B_processing_functions_nport.py` and `C_nport_build.py`): Python-based parsing and merging of N-PORT TSV files.
3. **Data Cleaning** (`C_clean_quarters.do`): Stata-based cleaning across quarters.
4. **ID Standardization** (`D_unique_report_holding_ids.do`): Creates consistent identifiers for holdings across reports.
5. **Master File Creation** (`E_masterfile_quarters.do`): Combines processed data into research-ready datasets.

3. One-Time Setup

Before running the pipeline, note that both the raw and the cleaned files will be saved in the folder containing the `NPORT_Master.do` file. Subfolders for each dataset will be automatically created.

3.1. Set Email in NPORT_Master.do

Before downloading the SEC N-PORT TSV files, include an existing email to request data. No account is needed.

```
global sec_email "your_email@institution.edu"
```

3.2. Specify Quarters in NPORT_Master.sh

Set the starting and ending quarters of the N-PORT panel dataset to be created.

```
local startq = "2019q4"
local endq   = "2024q4"
```

3.3. Confirm Stata Environment in NPORT_Master.sh

If your Stata environment requires user-contributed packages (egenmore, mmerge, ftools, gtools), simply uncomment the corresponding ssc install lines near the top of `NPORT_Master.do`.

```
// cap ssc install egenmore
// cap ssc install mmerge
// cap ssc install ftools
// cap ssc install gtools
```

3.4. Running the Pipeline

`NPORT_Master.do` controls the full build over the quarter range set. One can typically invoke it as:

```
do NPORT_Master.do
```

4. Outputs and Directory Structure

By default, all outputs are created under the folder that contains `NPORT_Master.do`. The pipeline creates the following subfolders automatically:

- `raw/` — SEC's TSV flat files as downloaded.
- `processed/` — Python-built DTA files with merged fund and security information.
- `cleaned/` — Stata-cleaned quarterly files.
- `extra/` — auxiliary lookups and mapping tables.
- `masterfile/` — research-ready panel dataset at the fund–security–quarter level.

5. Contents of the Quarterly Masterfile

The list below highlights the main variables used in the dataset and is not an exhaustive dictionary. Variable names are preserved exactly as in the SEC Form N-PORT schema, so please refer to the official N-PORT documentation for full definitions and additional fields.[1]

- ACCESSION_NUMBER – filing key.
- HOLDING_ID – position key within a filing.
- QUARTER_REPORT, QUARTER_REFERENCE – reporting quarter and filing quarter.
- REPORT_DATE, FILING_DATE, SUB_TYPE – timing and filing metadata.
- IDENTIFIER_ISIN, IDENTIFIER_TICKER, OTHER_IDENTIFIER, ASSET_CAT – security information.
- CURRENCY_VALUE, CURRENCY_CODE, EXCHANGE_RATE, INVESTMENT_COUNTRY – position information.
- ISSUER_NAME, ISSUER_TITLE, ISSUER_LEI, ISSUER_CUSIP, ISSUER_TYPE – issuer information.
- SERIES_NAME, SERIES_ID, SERIES_LEI, TOTAL_ASSETS, NET_ASSETS, CIK – fund information.

6. Citing and Contact

If you use the data or pipeline, please cite:

> Cavani, Maggiori, Schreger (2025), "GCAP Public Security-Level Data on U.S. Fund Holdings." Working Paper.

For questions or corrections: `info@globalcapitalallocation.com`.

References

Cavani, Bruno, Matteo Maggiori, and Jesse Schreger, "GCAP Public Security-Level Data on U.S. Fund Holdings," *Working Paper*, 2025.

[1] *Remark:* The compressed N-PORT archive is approximately 15 GB, and the uncompressed files total roughly 140 GB. Please plan storage accordingly.