HCP Data Collection Workflow Guide v4.2.3

Last Updated: September 6, 2025, 11:45 PM UTC

Purpose: Correct paths and data requirements for GARCH-ready collection

---

Overview

The HCP Data Collector uses a **hybrid approach**:

- **Automated**: Market data (Yahoo), economic indicators (FRED), IMF data (SDMX)
- **Manual**: Yuan SWIFT (PDFs), COFER USD, R&D/Revenue, Central Bank Gold
- **Master File**: Persistent storage of all historical data
- **GARCH Requirements**: 20 years (240 months) of data for robust volatility modeling

---

Directory Structure (CORRECTED)

```
C:/Users/markf/OneDrive/Documents/GitHub/Humble-Conviction-Portfolio/
└── data_collector/
    ├── hcp_unified_collector_v4.2.3.py
    ├── Outputs/
    │   ├── hcp_master_data.json (persistent data storage)
    │   ├── hcp_data_v423_*.json (collection outputs)
    │   └── hcp_manual_update_*.csv (for manual entry)
    ├── pdfs/
    │   └── [SWIFT RMB Tracker PDFs go here]
    └── logs/
        └── collector_*.log
```

Initial Setup (One-Time Process)

Step 1: Verify Directory Structure

Ensure the directories exist at the GitHub repo location:

```
bash
```

```
cd C:/Users/markf/OneDrive/Documents/GitHub/Humble-Conviction-Portfolio/data_collector
mkdir Outputs pdfs logs
```

Step 2: Download Historical PDFs

1. Go to SWIFT RMB Tracker
2. Download **ALL available monthly PDFs** (ideally 20+ years if available)
3. Save to: `C:/Users/markf/OneDrive/Documents/GitHub/Humble-Conviction-Portfolio/data_collector/pdfs/`
4. Naming doesn't matter - parser reads content, not filename

Step 3: Run Initial Collection (20 YEARS)

```
bash

python hcp_unified_collector_v4.2.3.py --initialize
```

This will:

- **Fetch 20 YEARS of market data** (for GARCH modeling)
- Parse all PDFs in the pdfs/ folder
- Query IMF SDMX API for COFER data
- Create `hcp_master_data.json` with all historical data
- Generate `hcp_manual_update_YYYYMM.csv` for missing data

Note: Initial collection takes ~5-10 minutes due to the 20-year data fetch.

Step 4: Add Missing Historical Data

1. Open `Outputs/hcp_manual_update_YYYYMM.csv`
2. Fill in missing columns:
   - **Yuan_SWIFT**: Should be populated from PDFs (if not, check PDF location)
   - **COFER_USD**: IMF reserves data (quarterly, % of global reserves in USD)
   - **RD_Revenue**: S&P 500 R&D spending as % of revenue
   - **CB_Gold**: Central bank gold purchases (tonnes, quarterly)
   - **US_Mkt_Cap**: US market cap as % of global (currently ~60%)

Data sources for manual entry:

- COFER: IMF COFER Database

- R&D/Revenue: FactSet, Bloomberg, or S&P Global
- Central Bank Gold: World Gold Council

Step 5: Import Manual Data

```
bash

python hcp_unified_collector_v4.2.3.py --import-csv hcp_manual_update_YYYYMM.csv
```

Result: Complete `hcp_master_data.json` with 20 years of historical data

Monthly Update Process

Step 1: Download New Month's PDF

1. Download latest SWIFT RMB Tracker PDF (released ~20th of each month)
2. Add to `data_collector/pdfs/` folder

Step 2: Run Monthly Update

```
bash

cd C:/Users/markf/OneDrive/Documents/GitHub/Humble-Conviction-Portfolio/data_collector
python hcp_unified_collector_v4.2.3.py --monthly
```

This will:

- Load existing `hcp_master_data.json`
- Fetch current market data (automatic)
- Parse new PDF for Yuan data (automatic)
- Check IMF API for new COFER data (automatic)
- Generate `Outputs/hcp_manual_update_YYYYMM.csv`

Step 3: Review and Complete Manual Data

Open the generated CSV file in `Outputs/`:

```
csv

Date,    DXY,   QQQ/SPY, Forward_PE, Yuan_SWIFT, COFER_USD, R&D_Revenue, CB_Gold
2025-09,  97.73, 0.890,   26.6,      ,          ,          ,
```

Fill in missing values:

- **Yuan_SWIFT**: Latest value from new PDF (should auto-populate)
- **COFER_USD**: Only if new quarter (Q3 2025 = July data, available September)
- **R&D_Revenue**: From quarterly earnings (if new quarter)
- **CB_Gold**: From World Gold Council (quarterly)
- **US_Mkt_Cap**: Update if significantly changed from 60%

Step 4: Import Updates

```
bash

python hcp_unified_collector_v4.2.3.py --import-csv hcp_manual_update_202509.csv
```

Step 5: Generate Tracker Input

```
bash

python hcp_unified_collector_v4.2.3.py --export-tracker
```

Creates `Outputs/hcp_tracker_input_YYYYMMDD.json` ready for the HCP Tracker.

Data Requirements for GARCH

Minimum Requirements:

- 60 months (5 years) - absolute minimum for basic GARCH
- 120 months (10 years) - recommended for stable estimates
- **240 months (20 years) - optimal for capturing multiple market cycles**

Why 20 Years?

- Captures multiple volatility regimes (dot-com, 2008, COVID, current)
- Provides robust parameter estimates
- Allows for out-of-sample testing
- Enables regime-switching GARCH models

Data Quality Checks: The collector reports "GARCH ready" indicators - aim for all 12 indicators to have 240+ monthly observations.

Quick Command Reference

```bash
# Navigate to correct directory first!
cd C:/Users/markf/OneDrive/Documents/GitHub/Humble-Conviction-Portfolio/data_collector

# First time setup (20 years)
python hcp_unified_collector_v4.2.3.py --initialize

# Monthly update
python hcp_unified_collector_v4.2.3.py --monthly

# Import manual data
python hcp_unified_collector_v4.2.3.py --import-csv hcp_manual_update_YYYYMM.csv

# Export for tracker
python hcp_unified_collector_v4.2.3.py --export-tracker

# Check current status
python hcp_unified_collector_v4.2.3.py --status
```

Troubleshooting

"No PDFs found"

Check that PDFs are in:

```
C:/Users/markf/OneDrive/Documents/GitHub/Humble-Conviction-Portfolio/data_collector/pdfs/
```

NOT in Desktop/pdfs/

"Master file not found"

Check that you're running from:

```
C:/Users/markf/OneDrive/Documents/GitHub/Humble-Conviction-Portfolio/data_collector/
```

And that Outputs/ directory exists.

"Not enough data for GARCH"

Run `--initialize` mode to fetch 20 years, not `--monthly` which only fetches 2 years.

IMF COFER Data Missing

- COFER is released quarterly with 1-quarter lag
- Q2 2025 data available in September 2025
- Check manually: IMF COFER

Monthly Checklist

- [ ] Navigate to correct directory (data_collector/)
- [ ] Download latest SWIFT RMB Tracker PDF to pdfs/
- [ ] Run `--monthly` command
- [ ] Review generated CSV in Outputs/
- [ ] Add any missing manual data
- [ ] Import CSV updates
- [ ] Export for tracker
- [ ] Verify data quality (check for 240+ months)

Data Retention Policy

- **Master file**: Never delete `hcp_master_data.json`
- **Backups**: Keep last 3 backups (auto-created)
- **Output JSONs**: Can delete older than 30 days
- **CSVs**: Can delete after importing
- **PDFs**: Keep all for re-parsing if needed

Notes

- **Initial Setup Time**: ~10 minutes for 20 years of data
- **Monthly Update Time**: ~5 minutes including manual entry
- **Storage Required**: ~50MB for complete dataset
- **Network Required**: Yes, for API calls
- **Python Dependencies**: yfinance, pandas, numpy, beautifulsoup4, pdfplumber, requests