IWH
Leibniz-Institut für
Wirtschaftsforschung Halle

CompNet

MDI Walkthrough:

The Environment & How to write Modules

Alessandro Zona Mattioli *
CompNet & VU Amsterdam

*** with input from Marco Miorandi***

3rd TSI Workshop,
Vienna, February 2024

This project has received funding from the European Commission; Directorate-general for Structural Reform Support under grant agreement No 101101853.

Index

- **Introduction**
- **Coding & Test Data**
  - Simulate test data
  - Replicate the condition of each country
- **Test code infrastructure**
  - Launcher routine
- **How to write your module**
  - General principles
  - Key functions
  - Embed in general code infrastructure

Introduction

How to approach the MDI?

- The CompNet Micro-Data-Infrastructure (MDI) is a powerful tool. How do we use it?

- I have an interesting analysis I would like to run code on cross country data, how do I proceed?

Meta Data

- The first step is to look at the list of variables available in the MDI:
  1. **Microprod_Metadata.xlsx**
  2. **MPnames_CCnames.csv** in each country

Meta Data (1)

- Once we are assured the data we need are available, we should check our sample sizes:
  1. **Overlap_firmid.xlsx**
     for each dataset, it will tell the overlap with the BR. The more datasets you wish to merge, the thinner the sample size is likely to get.

Detailed list of data types

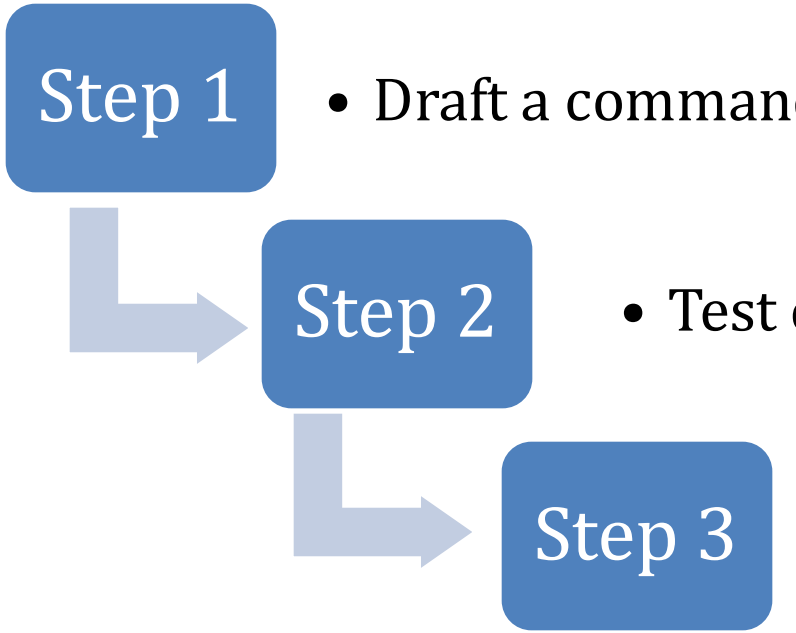
STEP 0

1. At the very beginning, make sure the data you may need **is available** in the MDI…
2. … and keep in mind the final sample size you will work with
   - Units of the data are key for merges

Coding & Test Data

Remote Coding

- We are used to develop codes as a trial-and-error process…



- **On the MDI, we can't follow this process, as we can't see the data.**

Solving the Issue

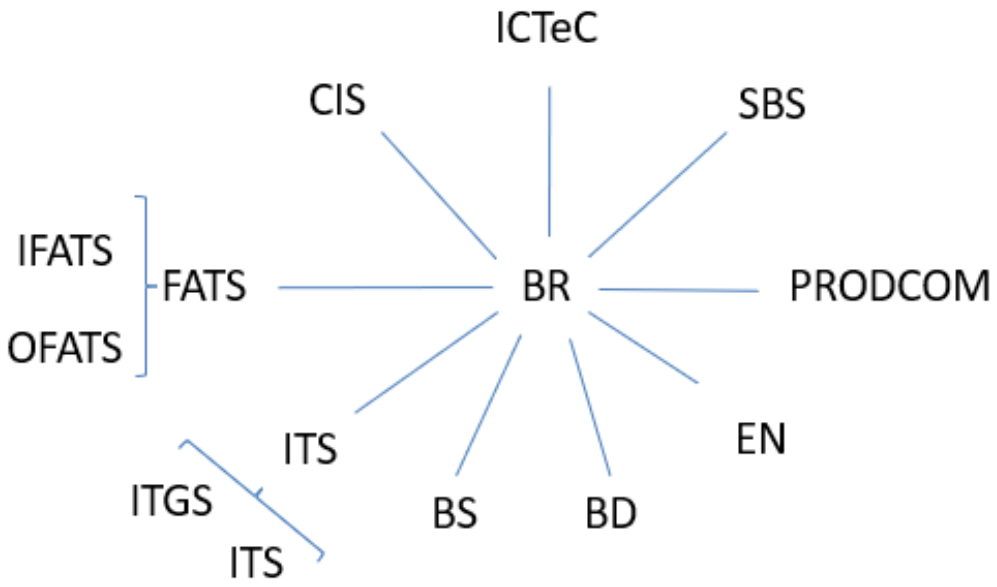
- We are working on providing Users with the resources to overcome this issue:

1. MetaData & Mock Data
2. Ready Made Functions
3. Self Guided Training

Mock Data

- A “synthetically" generated firm level dataset that allows to replicate the MDI data:

1. Create your own “scenario”
2. Adapt to the real condition in one of the MDI countries



Mock Data

- **Main Features**:
  - Mock data are **generated directly on your computer** via a replication package based on R syntax.
  - Come with a short **instruction manual** to understand how the data are generated and therefore what to expect from them.
  - They answer to the principle that the best way to learn about a dataset is to see it and work with it.

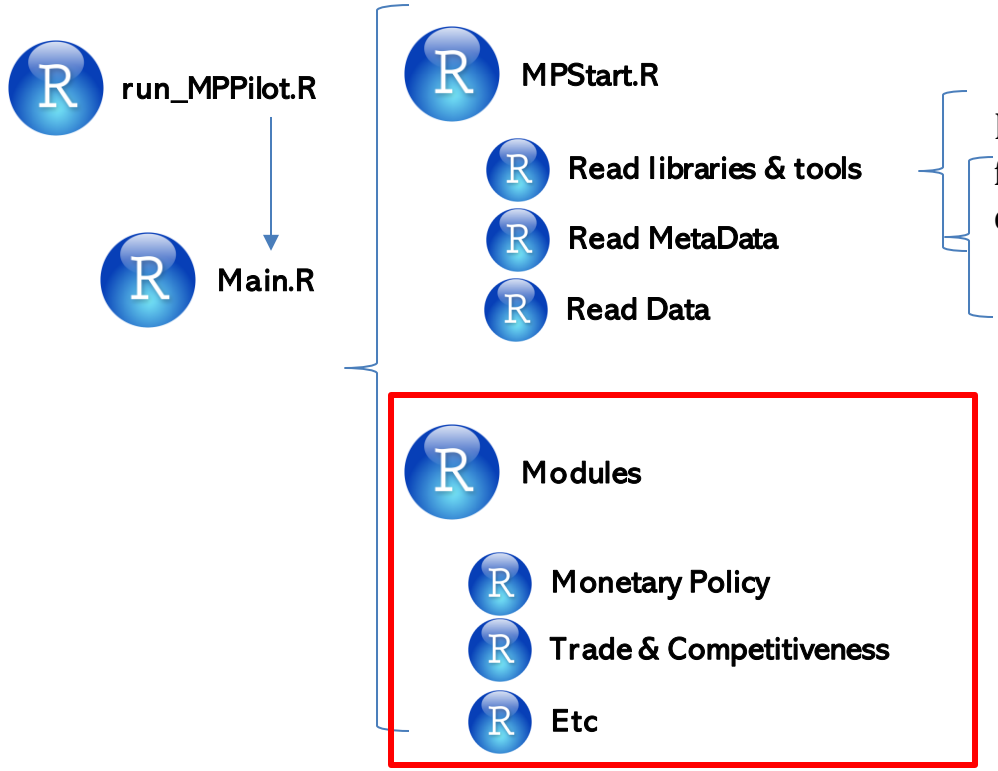
- The mock data **replicate the structure of the MDI in each country**:
  - Variable names
  - Files structure
  - Available indicators
  - Data Linking

- **Important to keep in mind**: a code that runs without errors on the mock data has very high chances of running without errors on the real data
  - Not all countries have 100% same data structure
  - Economic content of the mock data is necessarily simple.
- **Main consequence**: the same results are very unlikely to be obtained on the mock data and on the real data.

Test MDI code infrastructure

MDI is not just data…

- The MDI is also a **set of codes** based on **R data.table**.
- Designed to **periodically query subsections of the data** and run separate modules on them (“**rockets**”)



Advantages of this structure:

1. The code remains fixed from one rocket to the other.
2. Operations are standardized, reducing probability of bugs
3. Scalable to multiple data sources
4. Handy to manage (few files to update each time): only adjust the *console* once.

& metadata:

- **years & variables needed**

You can get acquainted with this yourself

- We provide a replication package that allows you to download the whole code package on your machine and test it on the mock data.
  - With some instructions on how to “install”
- Useful to master how to embed your modules in the main code.

How to Write Your Modules

MDI Modules

- Modules are the stage of the process where the analysis actually takes place.
- They should build on the previous code (MPStart)
- Users can write them in *"total freedom"*, a part from a set of limitations.
  - On some actions/operations
  - On the output

General Structure

1. On the header, indicate:
   - Which data & variable you are going to use & output you are going to create
2. Import data you need
   - Be mindful of memory!
3. Run your analysis
4. Export the results
5. Clean up!

Some DOs and DON'Ts

- **DOs**

1. **Conditional statements** to execute certain operations only in certain countries
2. Use our **tools** whenever possible
    1. Extract summary statistics
    2. Run regressions
    3. Other
3. Create few output files
4. Always run **disclosure routines**

- **DON'Ts**

1. Install packages or **download stuff from internet**
2. Load too much data
3. Merge allowing **cartesian products**
4. Create output without updating the Output description
5. Create output without including confidentiality parameters

Out tools

- General functions to execute common operations:
  - Cleaning data (trim/winsorize outliers)
  - Run regression and save output as a table
  - Extract summary statistics
  - Run structural estimations (e.g. production function estimation, perpetual inventory method)
  - Flag and blank non-disclosable cells

Out tools

- General functions to execute common operations:
  - Cleaning data (trim/winsorize outliers)
  - Run regression and save output as a table
  - **Extract summary statistics**
  - Run structural estimations (e.g. production function estimation, perpetual inventory method)
  - **Flag and blank non-disclosable cells**

How to extract summary statistics

1. In the data, provide a categorical variable that assigns a firm into categories (eg industries, technology classes, workforce composition, import mix, …)
2. If needed, allow for multiple hierarchical levels

```
KIC Description
TOTa Total Economy
HT high-tech industries
HTmfg -high-tech manufacturing
HTKIsv -high-tech knowledge intensive services
KI knowledge-intensive industries
MHmfg -medium-high tech manufacturing
KImsv -knowledge intensive market services (excluding high-tech and financial services)
KIfin -knowledge intensive financial services
KIoth other knowledge intensive services
Low other industries
MLmfg -medium-low tech manufacturing
LTmfg -low tech manufacturing
OTHmsv -less knowledge intensive market services
OTHsv -other less knowledge intensive services
```

```
h_0,h_1,h_2
HTmfg,HT,TOTa
MHmfg,KI,TOTa
MLmfg,Low,TOTa
LTmfg,Low,TOTa
KImsv,KI,TOTa
HTKIsv,HT,TOTa
KIfin,KI,TOTa
KIoth,KI,TOTa
OTHmsv,Low,TOTa
OTHsv,Low,TOTa
```

3. Call the function on the target dataset, indicating:
    1. Which summary stats you need
    2. and on which variables

Flag and blank non-disclosable cells

- Exporting output files
- Be mindful of the fact that someone will need to inspect that the output files are disclosure-free
  - Provide description
  - Indicate # obs underlying each statistics & dominance parameter
  - Do not generate too large files

- The first 2 steps are taken care of in our export_db function
  - Automatically updates an OutputDescription.txt file
  - Automatically blanks bad cells based on the country specific requirements

Documentation

- We are preparing a tool library to facilitate use of our functions

- The more you rely on them, the lower the probability of running into errors, breaking the code and have to spend time debugging

Conclusions

In short, steps to follow:

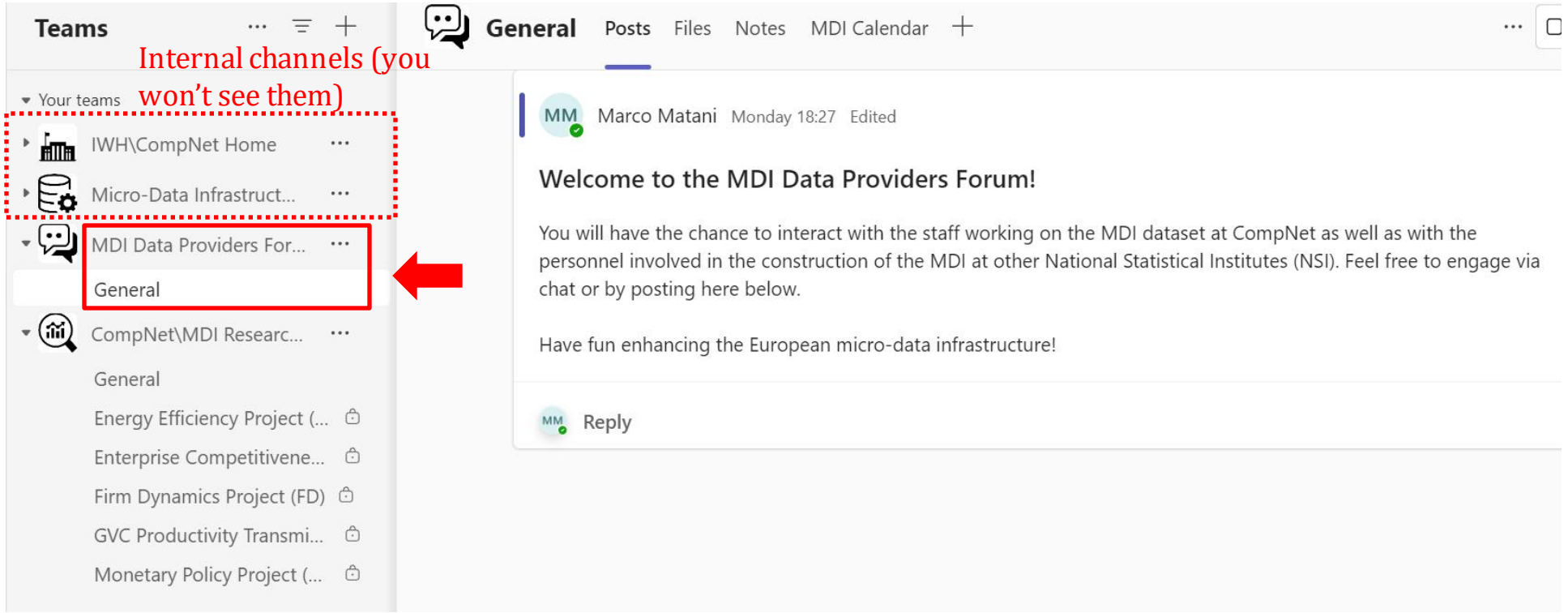
- Step 0: check the metadata
  - Assess where your desired variables are available
  - Consider the possible sample size
- Step 1: draft code on the mock data
  - Start by playing around and then draft a module
- Step 2: send module and required metadata to the MDI team in time for the next *rocket launch*
- Step 3: take a deep breath and get ready for some debugging…

New Teams Infrastructure

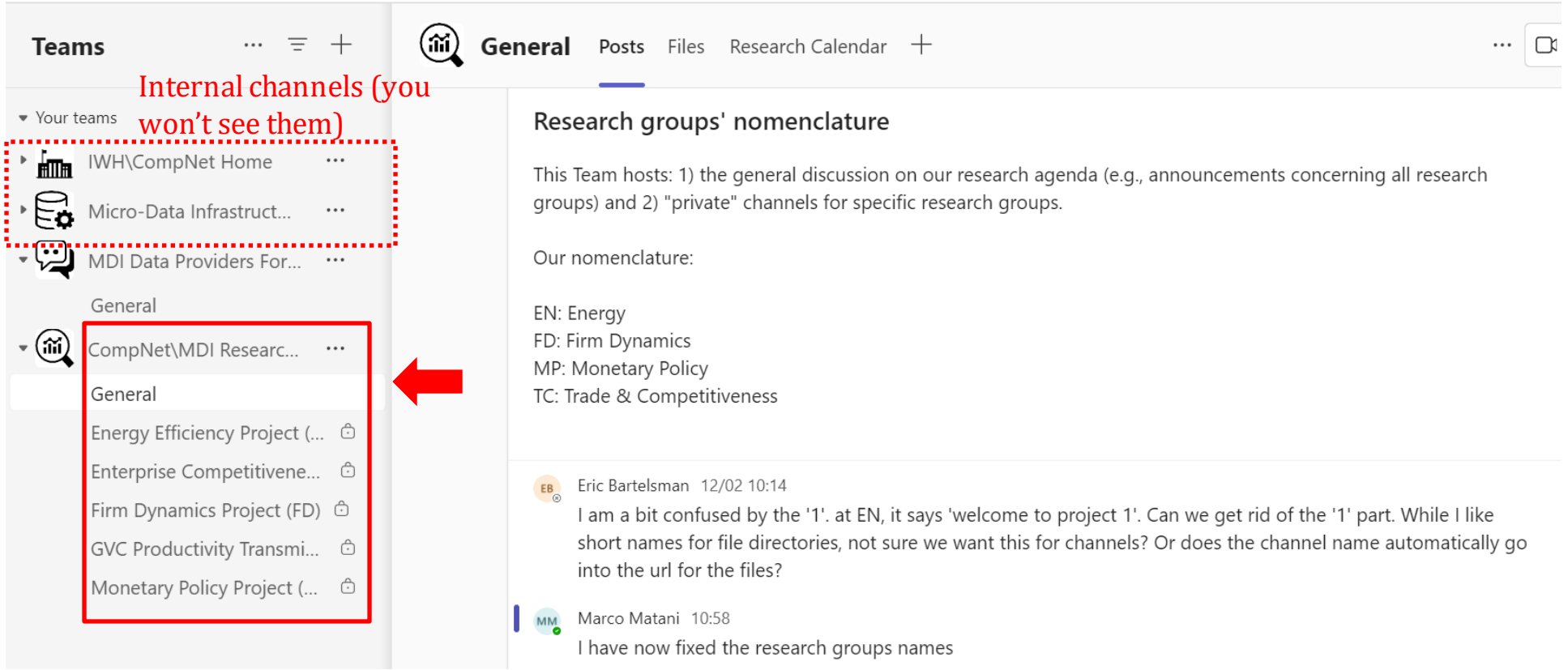
Purpose

- Meet the needs of NSIs in the MDI to have a “data provider forum”
- Facilitate interactions between CompNet staff and external collaborators (both NSI and NPB)
- Promote the exchange of best practices in constructing the MDI
- Foster more seamless collaboration on common research projects with NPBs
- Slash down the amount of exchanged e-mails

Structure (for MDI/NSIs)



Structure (for Research/NPBs)



Discussion

Thank you!

Any Questions?

Appendix

Type of data based on coverage

Data Source	Type of Data
BR	Census
SBS	Survey/Census
ITGS	Census
ITS	Census
CIS	Survey
ICTeC	Survey
OFATS	Census
IFATS	Census
BD	Census
BS	Survey/Census
ENER	Survey
PRODCOM	Survey
GVC	Survey