CompNet The Competitiveness Research Network

MDI data construction: Lessons learnt from the CompNet / MDI staff side

Marco Miorandi
Marcelo P Ribeiro *

(*) with input from Eric Bartelsman

February 23, 2024

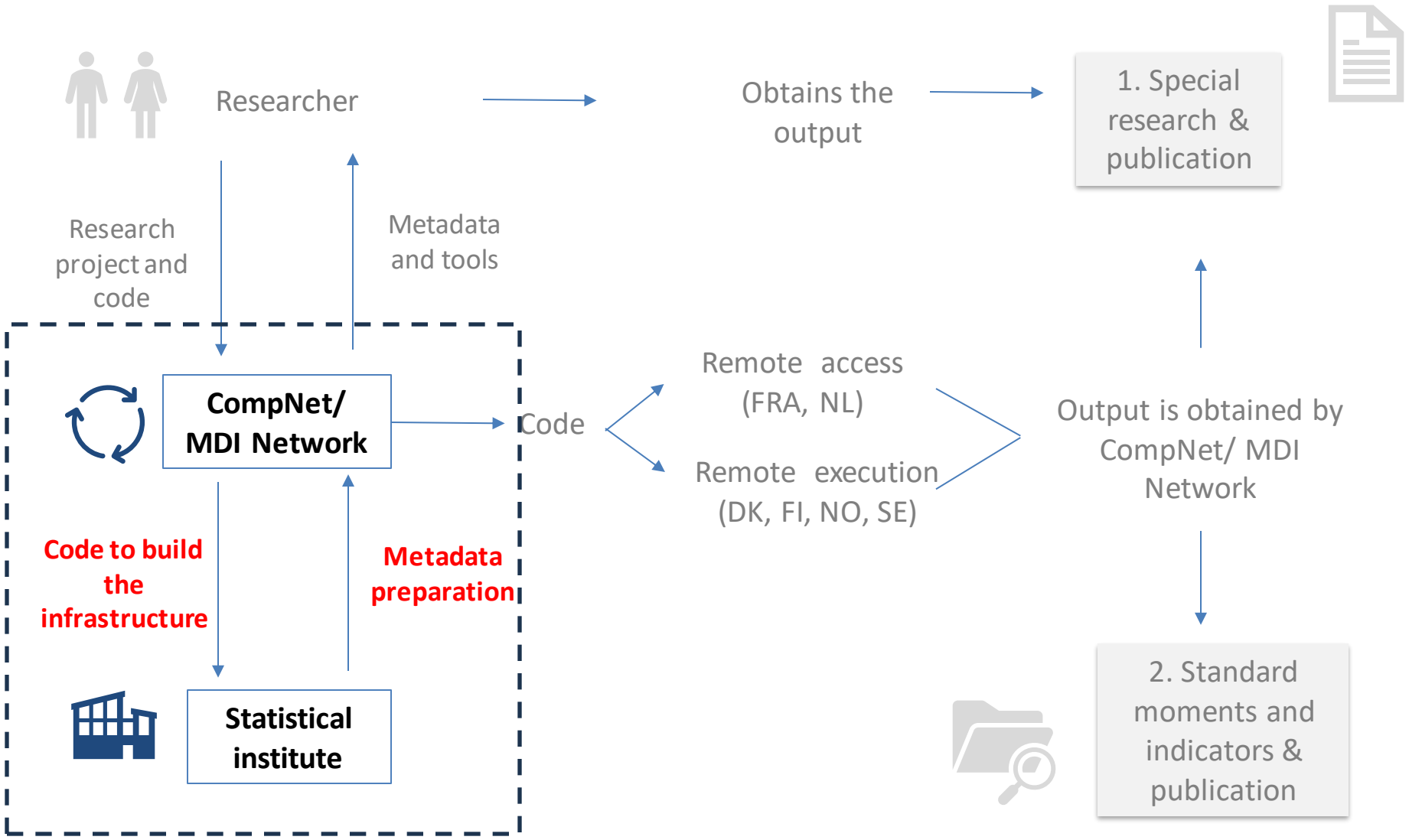
Introduction: MDI data construction

Goal: Constructing MDI data from NSI raw data

- Data construction process should
  - Be flexible to raw data from different NSIs
  - Require as little periodic rewriting of files

→ For this, **metadata of the raw data are key**

MDI architecture



Metadata preparation

- This part of the data construction process is needed to collect as much information on all the available raw data
- The corresponding metadata files should be collected and updated periodically
- This information has to be **provided by the NSI**

Metadata preparation

1. Define whole set of (raw) data files available

-> Create file '*NSI*_datafiles.csv'

For Example:

dataset	fileName	year	datasetType	location	otherCharacts
BR	busreg_08.csv	2008	csv	/C:Data/...	'business register data for sole proprietors'

2. Dataset-specific metadata

-> Create file '*NSI_dataset*_varlist.csv'

For example, for dataset 'busreg_08.csv':

var	descr	domain	units
tot_assets	'Total assets owned in a year'	value	euro
nace	'NACE industry code'	character	code

3. Classification lists

-> Create file '*NSI_classification*_class.csv'

For example, for classification variable NACE:

nace	year
0001	2009
0002	2009
0001	2011
0007	2013

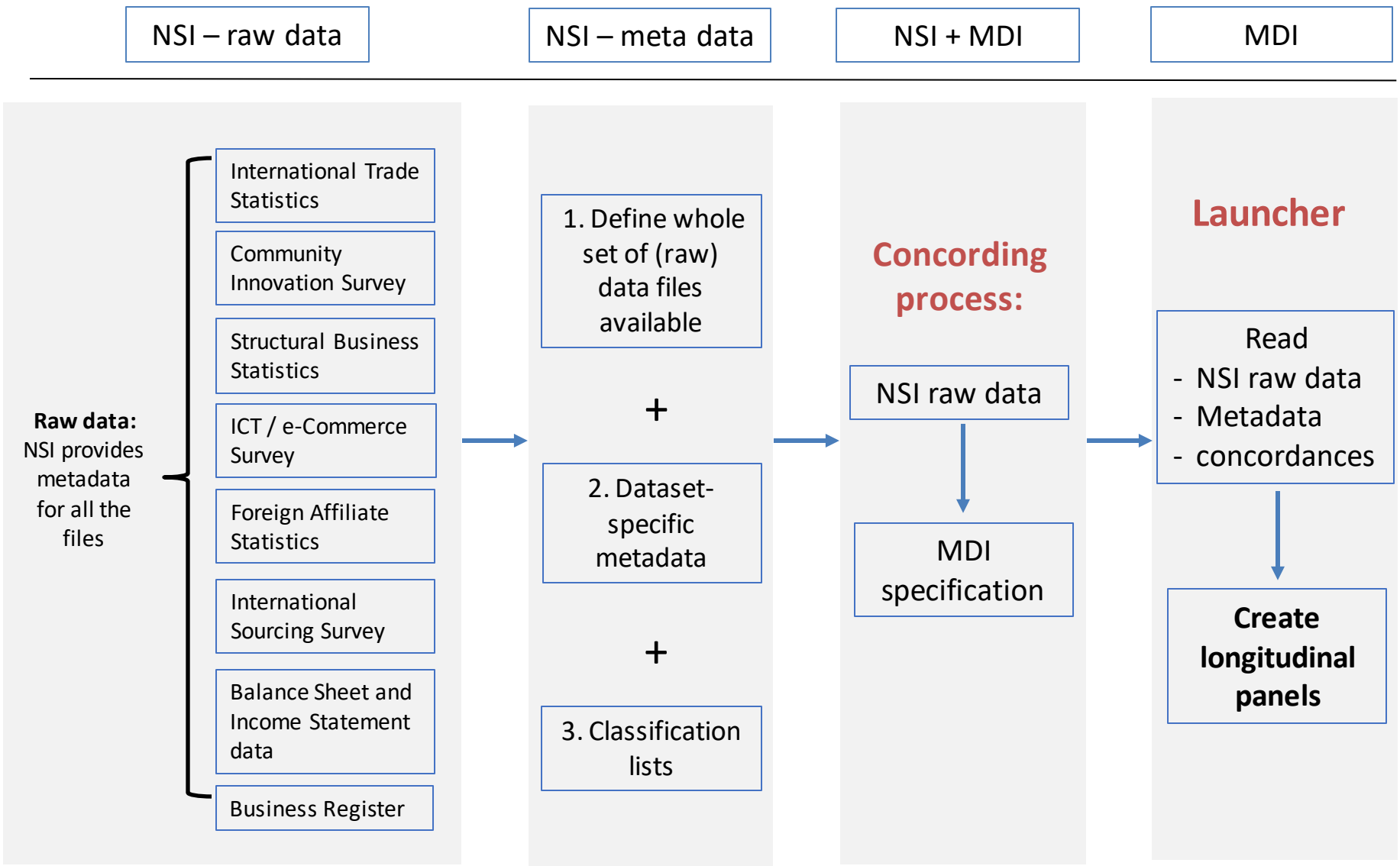
Harmonization process

- This part of the process requires to concord through rows, columns and values
- The three metadata files previously defined are **essential** for this operation
- Then **concordances** for files, variables, records (rows), and classifications are **created by the NSI and us together**

→ **Key row-identifying variable**: '*firmid*'

--> then, use BR to map the firm to other datasets

From raw to MDI data



From raw to MDI data: Example

Community Innovation Survey ("RAW")

Cis_2001.csv →

ID	yyyy	capital_int	SInace2001
32112	2001	389,000	2340

Cis_2002.csv →

firmID	Year	K_int	SInace2002
32112	2001	381.000	2340

Cis_2003.csv →

firmID	Year	Phys_K_int	SInace2003
32112	2001	681.000	23401

Cis_2004.csv →

ID	Year	Phys_K_int	SInace2004
32112	2001	681.000	23401

Things to harmonize

1. Columns:
   - *ID-firmID*
   - *yyyy-Year*
   - *capital_int-K_int*
2. Units of observation:
   - legal unit, a firm, plant…
3. Classification:
   - Are classifications the same across the years (e.g., Nace Rev 2)

From raw to MDI data: Example metadata

Harmonization depends on a rich metadata, i.e :

Variable	Data Source	Label/Description	Format	Values
ENT_ID	BR	Unique enterprise identification	Character	X
ENTgrp_ID	BR	Enterprise Group ID	Character	X
Start_Ent	BR	Start date for the enterprise ID	Character	Date DDMMYYYY
End_ent	BR	End date for the enterprise ID	Character	Date DDMMYYYY
LEGAL	BR	Legal form of the enterprise ID	Character	LL= Limited liability company - include limited liability partnerships and public corporations, SP= Sole Proprietor ND= Not defined

- For each variable: what is the variable domain (boolean, character, etc.)

- Rich description allow us to create formula to concorde, e.g.,
  - 62 variables related to investment in AT
  - 32 variables related to investment in IT
  - Formula create MDI investment variables (common/equivalent to AT and IT)

--> Use

From raw to MDI data: Example

Harmonization = Mapping file

	A	B	C	D	E	F	Q	R	AA	AB	AC	AD
1	MPname	DataSource	Description	var_2000	var_2001	var_2002	var_2013		Values	uniqueDim	indexDim	Type
2	firmid	sbs	Research ID number (transformation of: Res	MS_POD_r	MS_POD_r	MS_POD_r	MS_POD_razST	...		1	1	
3	year	sbs	Year	LETO	LETO	LETO	LETO			0	0	
4	persons	sbs	Persons in paid employment by enterprise	ST_ZAPOS	ST_ZAPOS	ST_ZAPOS	ST_ZAPOS_POD			0	0	
5	month	sbs	Month	MESEC	MESEC	MESEC	MESEC			0	0	
6	personid	sbs	Statistical identification for person	SID_razST	SID_razST	SID_razST	SID_razST			1	1	
7												
8												
9												

	Harmonization	Solution
1	Same variable changes its name from year to year	Map the name of each variable to MDI equivalent name Ex: Var_2000 and var_2001 could have different names in the raw csvs but they are being called as the MPname now
2	Possible breaks: Do *capital_int* and *K_int* means the same a s *Phys_K_int*?	Concordance formulas: for example, in SI all the types of investment are from a hierarchical classification (ie Total, tangible, machines, drill press). Traverse to find commonality across counties and years.

From raw to MDI data: Example

Community Innovation Survey (RAW)

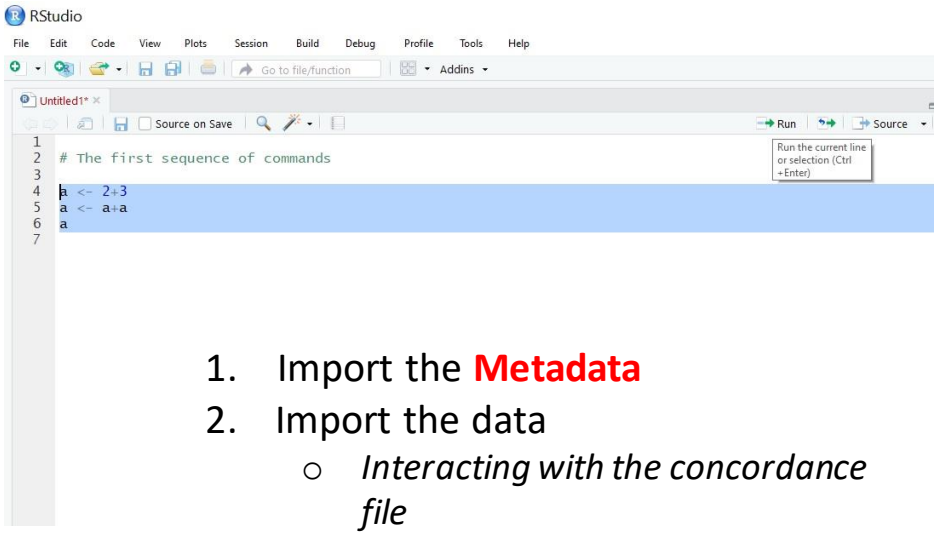
Cis_2001.csv →

Cis_2002.csv →

Cis_2003.csv →

Cis_2004.csv →

R script: *MPStart*



1. Import the **Metadata**
2. Import the data
   - *Interacting with the concordance file*
3. Creates a panel (firmid-year)
4. Temporary save one single file: *CIS.rds*

Wrapping up

- The data construction process in steps

1. Obtain raw data files from NSI
2. Obtain metadata on raw data from NSI
    1. List of files
    2. List of variables
    3. Classification lists
3. NSI/MDI concordances between NSI and MDI specification
4. MDI launcher code
    - ➢ read raw data, metadata, concordances and create harmonised longitudinal panels.

Thanks for your attention!

Any questions?