WORLD BANK GROUP
Development Economics Data Group

Best practices and automated tools for industry classification

Experience from the World Bank

by Arthur Giesberts (AGiesberts@worldbank.org)

Acknowledgement
All achievements are based on a strong and pleasant cooperation with National Statistical Institutes in many countries and on the efforts and commitment of my colleague Shwetha Eapen.

Group of Experts on Business Registers
26-29 September 2022

Session 1: Classifications and identifiers in the SBR

Business Statistics at the World Bank

The World Bank supports the development of Business Statistics for a range of applications, including measuring innovation, productivity and employment through:

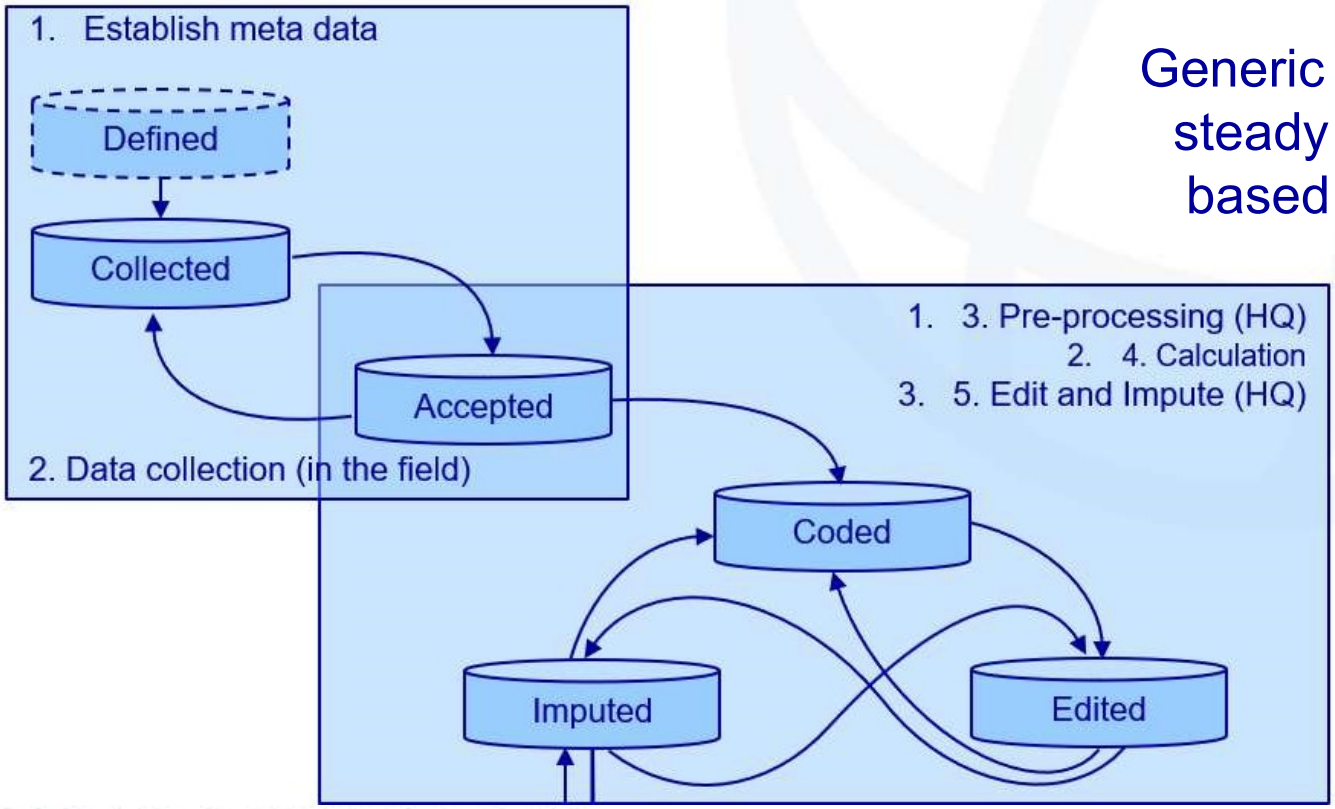
- Statistical capacity building with National Statistical Institutes of client countries
- Financing economic censuses and surveys
- Providing technical assistance (TA) for economic censuses, surveys and the maintenance of business registers.
  - Design of lean (smart, cost-effective and sustainable) statistical processes
  - Design of survey instruments (like questionnaires, training materials)
  - Development of **streamlined statistical tools** for sampling, data collection, industry classification, editing and imputation.

- This presentation focuses on TA activities, in particular a streamlined statistical tool implemented for **industry coding**.

Technical Assistance in Business Statistics Processes

The industry classification tool forms part of:

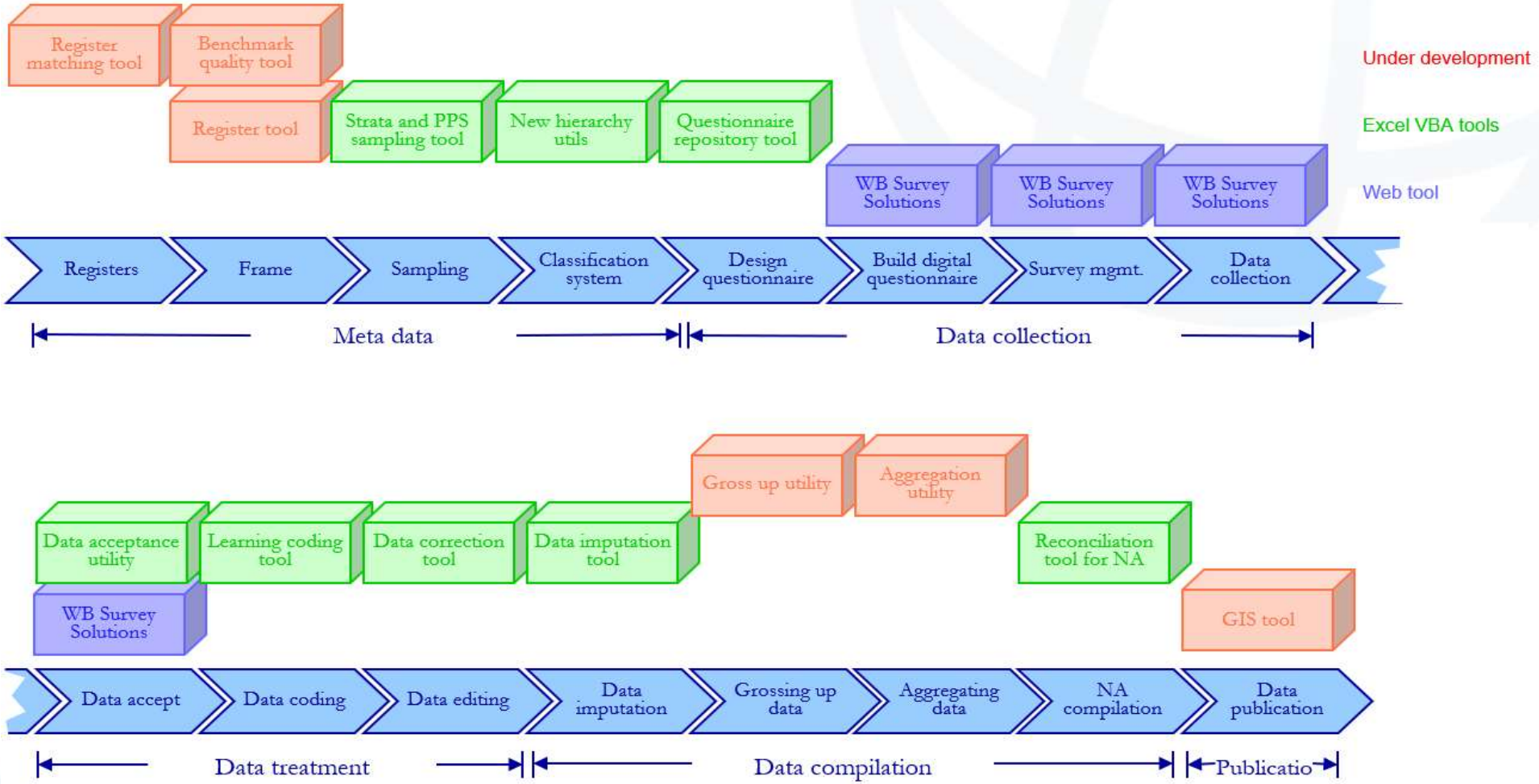
- A broader World Bank work program to improve business statistics, as infrastructure for national accounts and other indicators
- Lean design of statistical processes



Generic process based on steady stages and rule-based transformations

Technical Assistance in Business Statistics Tools

A toolbox of standard methods and applications to support data collection and compilation in business statistics



WORLD BANK GROUP
Development Economics Data Group

Industry coding for economic statistics

- High quality business statistics and national accounts require accurate, comprehensive and timely industry classifications that adhere to official standards.
- Industry coding is a significant task requiring substantial knowledge, experience and resources.
- In practice, coding often relies on individual practices.
- Explicit rules can be developed and automated for generalized application in:
  - Support for manual coding
  - Fully automated coding
  - Validating available codes in dataset to:
    - Identify structural errors
    - Assess overall quality level

Benefits and challenges

While automation brings many potential benefits:

- Objective rules
- Timely, efficient processing
- Reproducible results
- Ruleset reusable over time
- Traceable coding process
- Self documenting, including para data

It comes with a number of challenges and must address:

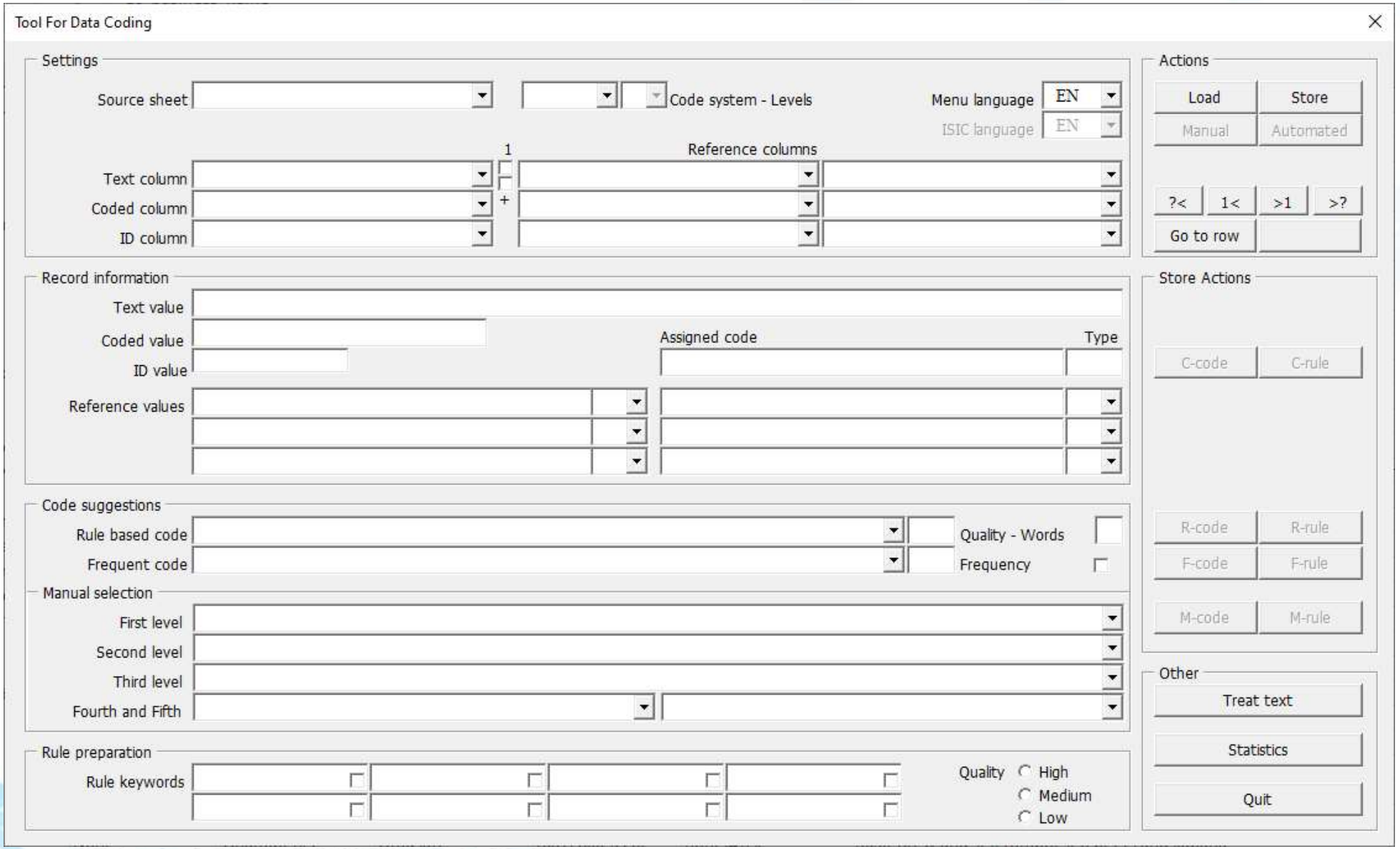
- Multiple languages
- Poor data quality (from inadequate questionnaire design staff training)
- Spelling errors
- Correct/incorrect use of identical words in different contexts
- Typical national or regional terminology

Automated tools must be simple and transparent for use in a wide range of local circumstances.

Key features of the tool

- Requires variables like main activity, name of the company, main product or similar, or a combination of these
- User-friendly multilingual VBA interface – applied in over 10 languages including, for example, Arabic and Khmer
- Can be used in fully automated mode or facilitate manual coding, to validate existing classifications or to classify secondary activities.
- Combines the benefits of:
  - MS Excel (WYSIWYG and filtering)
  - A standard software program reducing the risk of programming error.
- Generates para data and a log to monitor and evaluate coding process
- Houses an extended ruleset - new rules can be added or existing rules adjusted
- Can account for typical national economic activities or local terminology and address specific classification challenges/nomenclature
- Usage is scalable for large datasets - tasks can be divided effectively

Interface of industry classification tool



WORLD BANK GROUP
Development Economics Data Group

Methodology

The tool uses a combination of steps and rulesets to facilitate the industrial classification process (either manual or automated):

Basic mode

1. Remove “trivials”: Unnecessary distorting characters are removed
2. Enlarge “shorts”: Since the tool requires at least 4 letter keywords
3. Correct “spellings”: Correction for spelling errors
4. Interpret “singles”: For single words with insufficient context
5. Redirect “channels”: Convert to default keywords prior to coding
6. Apply coding rules: Linking 4 letter keywords to ISIC codes

Advanced mode

Additional variables in the dataset can be used to refine industry assignment (for example, cost structure).

Widely used in World Bank projects

Industry classification tool used for approximately 50 countries in support of a range of statistical projects

Country examples 1/2

Country 1 (West Africa)

Census and economic survey of approximately 90,000 units

- Manual classification led to implausible estimates
- Using the industry coding tool, out of 88,258 records:
  - 50,545 out of 80,529 automatically coded records were assigned the same 4-digit ISIC code as manual
  - For the remainder, a comparison was made at the 1-digit level and where differences were identified, the largest firms were corrected manually

Country 2 (Middle East)

Administrative data and survey-based SBR of over a million records

- Inconsistent activity descriptions or codes across administrative data sources
  - Each record classified 3 times based on separate variables/registers
  - Algorithm used for final classification
- The first ISIC code listed often erroneously selected (while other variables automatically coded indicated another code), which occurred from confrontation

Illustrative example

Original ISIC \ Tool assigned ISIC	Blank	A	B	C	D	E	F	G	H	I	J	K	L	M	N	P	Q	R	S	Total
A	3	19		9		1	16	17	5	2					4	1				77
B			15	7	2		9													33
C	33		4	665	2	1	75	49	2	6	2		1	1	10	1	1		1	854
D				3	1		2													6
E				3		15	9								3					30
F	23	1	3	26	1	8	951	12	4	7	2		3	3	35	2			3	1084
G	18	2	1	34		1	37	431	13	5	1	2	2	4	3		1	1	2	558
H	8		1	4		2	14	8	184	1		1		1	5			1		230
I	9			1			12	12	3	245		1	2		3		1	1		290
J	2			1			12	3			41			1	4				1	65
K	3						2	2				58	1							66
L	5						19					1	15	1	2					43
M	9			7	1	1	50	8	1	1	3	3		101	3	1			1	190
N	2	1	1	5		3	45	4	9		2			2	140					214
O							1									3	1			5
P	9			1			5				2			1	1	180	1			200
Q	9							6		1				2	1	1	196			216
R				1			1			2								10		14
S	2			1			1	4		1	2				1			1	7	20
U														1						1
Total	135	23	25	768	7	32	1261	556	221	271	55	66	24	118	215	189	201	14	15	4196

WORLD BANK GROUP
Development Economics Data Group

Country examples 2/2

Country 3 (Asia)

Census and economic survey of 700,000 entities

- Classification tool allowed for automated validation of manually assigned codes, even though the appropriate (main activity) was not digitalized:
  - 68% identical at 1 digit ISIC
  - 18% could not be coded (since appropriate variable was not digitalized)
  - 10% of errors limited to two ISIC section levels

Country 4 (West Africa)

Census and economic survey of over 1,100,000 entities

- Out of over a million records about 10 thousand were corrected or assigned an ISIC code
  - Many typical errors corrected (pharmacies, tailoring services or repairs)
  - Significant number of spelling errors could be easily resolved (phonetic spelling)

Examples of lessons learned

- The terms "*main activity*" and "*secondary activity*" not clearly understood by respondents.
- Description of main activity should be collected and digitalized. If only codes are captured in the collection process, results cannot be verified.
- Typical errors can be systematically identified and corrected.
- Offsetting errors can be measured, and their impact determined.
- Variables easily collected can enhance the quality of results:
  - Name
  - Urban/Rural or GPS
  - Production/Trade/Services (Repairs)
  - B2B/B2C/B2All/B2GOV
  - Cost of goods bought for resale (y/n of value)

  Examples:
    - "Rice" → agriculture, production or trade (+ retail or wholesale)
    - "Water" → production (+ like) or trade (+ retail or wholesale)
    - "Computers" → production, trade or repairs

Summary

Implementation at the World Bank has demonstrated the automation of industry coding can:

- Result in a versatile, efficient and effective tool of benefit to national statistical offices
- Lead to significant improvements in cost-effectiveness and data quality
- Form an essential part of the iterative process of data compilation
- Integrate effectively with other Business Statistics tools within the overall statistical process, such as tools developed for data editing and data imputation

THANK YOU
Questions?