

EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS 2024 ESTORIL - PORTUGAL

Assessing Fitness for Integration - A Metadatadriven Approach

Thomas Gottron, Andrea Novello, Ilias Aarab, Bernadette Lauro European Central Bank





The conference is partly financed by the European Union



Assessing Fitness for Integration

A Metadata-driven Approach



06/06/2024

<u>Thomas Gottron</u>, Andrea Novello, Ilias Aarab, Bernadette Lauro European Central Bank

The challenge



Data analytics is often based on **combining granular datasets**.

Users start by addressing **three questions**:

- (1) Which data sets can be integrated?
- (2) How to perform integration in a semantically meaningful way?
- (3) How good is the linkage rate?

Idea: Provide a dashboard to answer the questions



Information on datasets, their structure and where to find them Integration of actual data, based on metadata information Standardised metrics to assess how well datasets can be linked and integrated Self service dashboard to explore datasets and integration metrics

Fully automated process

How does it work - Step 1: Metadata

Metadata



Identifier				
Time				
Unit				
Stratification				

- **Integrability** *I*(*A*, *B*): indicates if two datasets can be integrated at all, based on common **identifiers**.
- Linkage rate $L_i(A, B)$: ratio of identifier values present in A that can also be found in B, considering time information, e.g. reference period, validity ranges.
- Weighted linkage rate $L_{u,f}(A, B)$: based on a measurable **unit**, e.g. reflecting linkage weighted by market values, outstanding nominal amount etc.
- **Breakdowns: stratification** attributes are used to analyse integration on subsets of data.

How does it work - Step 2: Determine integrability





ID	U	S	T _A	ID	Τ _B

www.ecb.europa.eu ©

How does it work – Step 3: Compute metrics



How does it work – Step 4: Visualise



Prototype dashboard – Overview of datasets



Note: due to confidentiality of some datasets, the visualisations are based on artificial data

Prototype dashboard – Details for a single dataset

Integration [Dashboard	=					Î
Search		Datase	t Description	Use	ful Links		
Datasets for Integr Datasets Metrics AnaCredit CSDB	Time Pla The time	ot Stratif	ication Plot				
GLEIF MMSR ORBIS RIAD SUBA	134014 134014	'The Str 12,000	Country Map The map shows the number of units (e.g. entities) per cou	ntry.			
29 SUBA Dataset Integratio	134014 9 134013 134013 134013	9,000	T Kiribati Samoa	Cree Canada United States of America Mexico Cuba Colombia Peru Brazil Argentina Chile	Inland Iceland Norway France Ukraine Kazakhstan Gellen Italy Uzbekistan Algeria Sayria Algeria Saudi Arabia India Mali Chad Ethiopia Democratic Republic of the Congo Mozambique South Africa	Russia Mongolia China Japan Philippines Marshall Islands Indonesia Solomon Islands Fiji Australia New Zealand	
		-		0	500k 1000k 15		Highcharts.com © Natural Ea
ntiality of so	ome data	asets,		10			www.ecb.europa.eu ©

Note: due to confidentiality of some datasets, the visualisations are based on artificial data

Prototype dashboard – Integration of two datasets

Integration Dashboard	=	
Search Q		Select dataset to integrate
		RIAD
Datasets for Integration	Select source table to integrate	Select target table to integrate
Datasets Metrics <	crp_anacredit_anacredit_dmc_entty_cc	crp_riad_n_essential.riad_entty_flttnd_essnttl_d
Dataset Integration Metrics <		
» AnaCredit	Select the unit to integrate on	
	entity	
>> GLEIF	Run	
	Country Map	Country Barplot
	The map shows the number of units (e.g. entities) per country when the two datasets are integrated	The barplot shows the number of units (e.g. entities) per country when the two datasets are integrated
Alternative Integration	accordingly.	accordingly. Only the top 20 are shown. 'Other' shows the sum of the number of units of the countries not in the top 20.
	+ -	3,000,000
	Iceland Durate	
	Canada Norway Kussua France Ukraine Kazakhstan	
	United States of America Spain Greece Uzbekistan Algeria Iran China Japan	2,000,000
	MEXICO Cuba Mali Saudi Arabia india Philippines Grenada Guinea Ethiopia Marshall Islan	
	moa Peru Brazil Mozambique Fiji	
	Australia Argentina South Africa	1,000,000-
	Chile	

Note: due to confidentiality of some datasets, the visualisations are based on artificial data 11

Fitness for Integration



- The Fitness for Integration dashboard **supports users** in common data integration and exploration tasks
- The report is **metadata driven and automated** to reduce the need for manual investigations
- A prototype has been implemented and tested with users

Thank you and happy to take questions 😳



EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS 2024 ESTORIL - PORTUGAL



INSTITUTO NACIONAL DE ESTATÍSTICA



The conference is partly financed by the European Union