EUROPEAN CONFERENCE ON QUALITY IN OFFICIAL STATISTICS
2024 ESTORIL - PORTUGAL

Assessing Fitness for Integration - A Metadata-driven Approach

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

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

eurostat

The conference is partly financed by the European Union

EUROPEAN CENTRAL BANK

EUROSYSTEM

Assessing Fitness for Integration

A Metadata-driven Approach

06/06/2024

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

The challenge

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

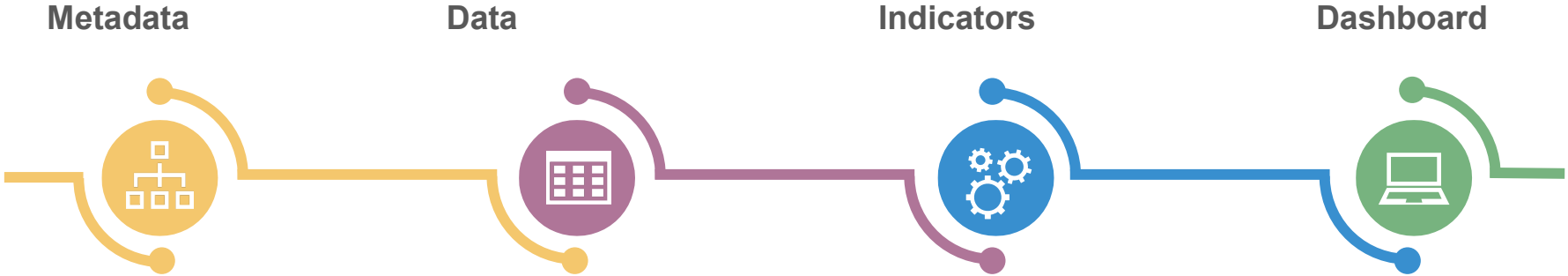
Users start by addressing **three questions**:

(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

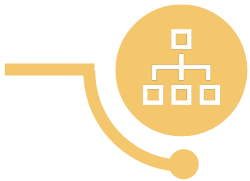
Standardised metrics to assess how well datasets can be linked and integrated

Self service dashboard to explore datasets and integration metrics

Fully automated process

How does it work – Step 1: Metadata

Metadata

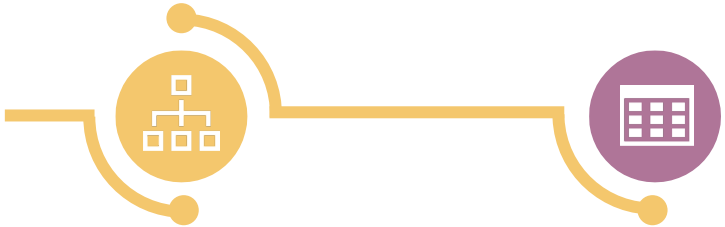


- **Integrability** $I(A,B)$: indicates if two datasets can be integrated at all, based on common **identifiers**.
- **Linkage rate** $L_i(A,B)$: ratio of **identifier** values present in A that can also be found in B, considering **time** information, e.g. reference period, validity ranges.
- **Weighted 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

Metadata

Data



- Identifier
- Time
- Unit
- Stratification
- …

Dataset A

ID	U	S	T_A

⋈

Dataset B

ID	T_B

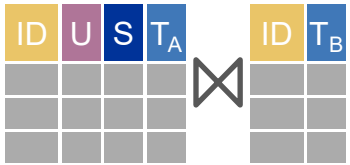
ID	U	S	T_A	ID	T_B

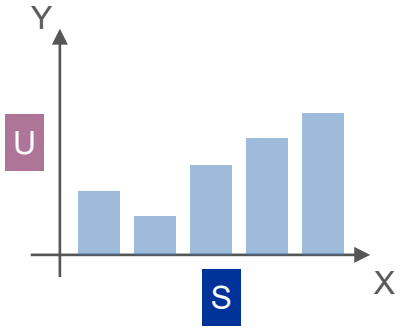
How does it work – Step 3: Compute metrics

Metadata **Data** **Indicators**

Source	Target	Metric	U	S	T
Table A	Table B	Linkage rate	Y_1	X_1	01/23
Table A	Table B	Linkage rate	Y_2	X_2	01/23
Table A	Table B	Linkage rate	Y_3	X_3	01/23

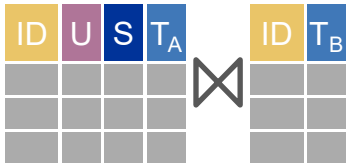
- Identifier
- Time
- Unit
- Stratification
- …

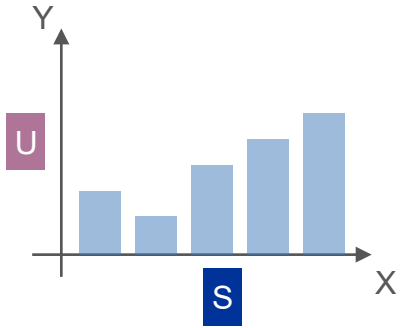


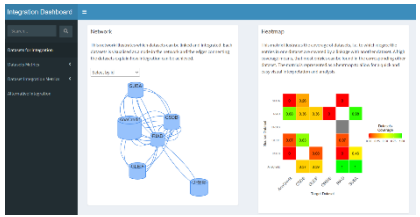


How does it work – Step 4: Visualise

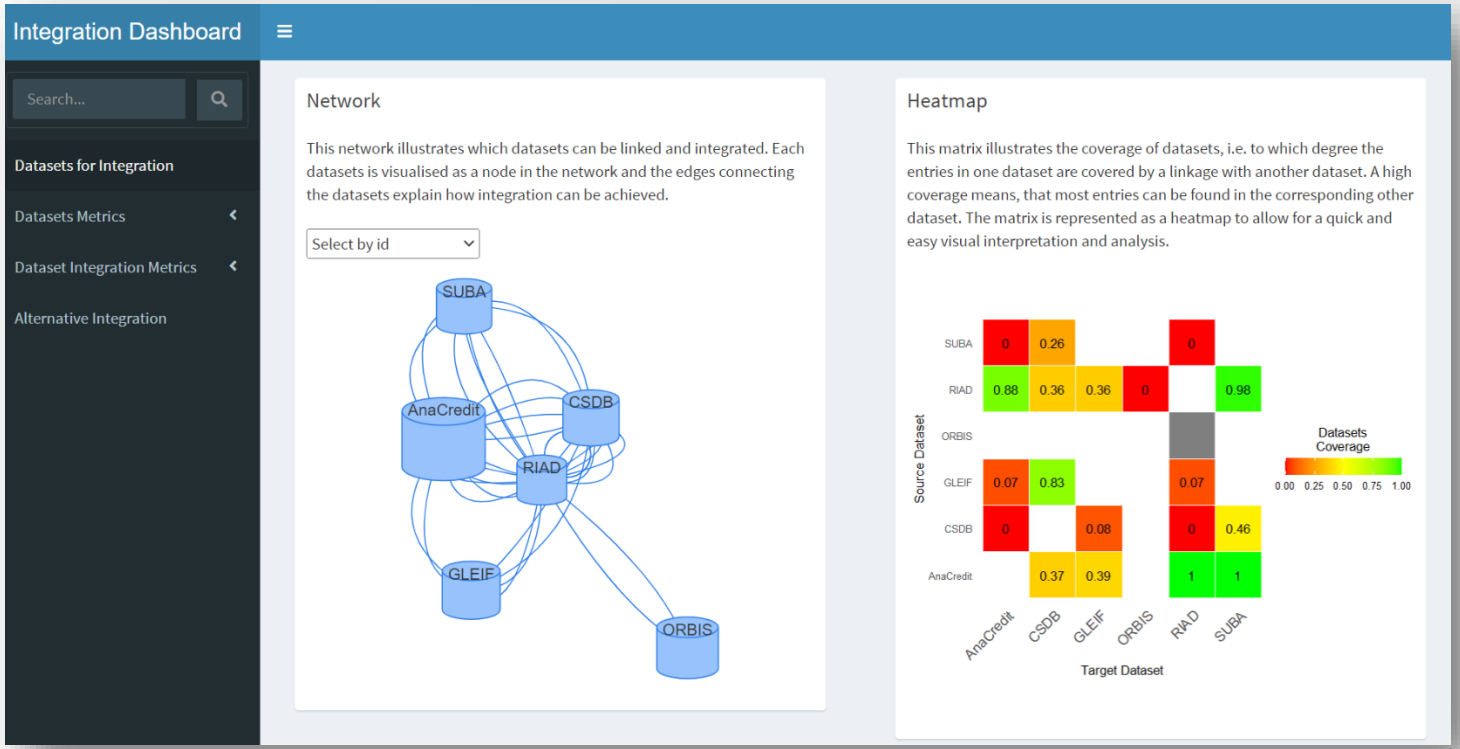
Metadata **Data** **Indicators** **Dashboard**





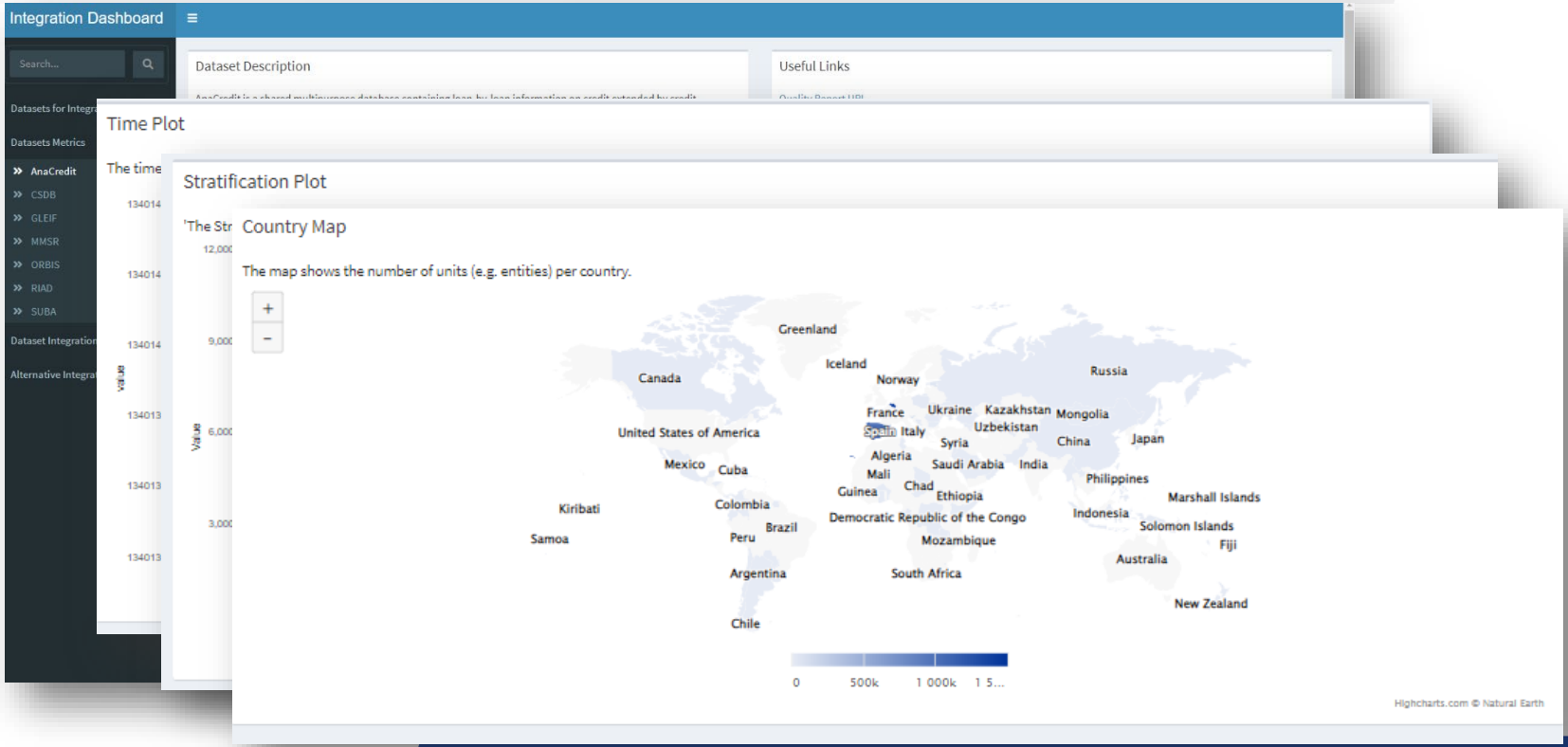


Prototype dashboard – Overview of datasets



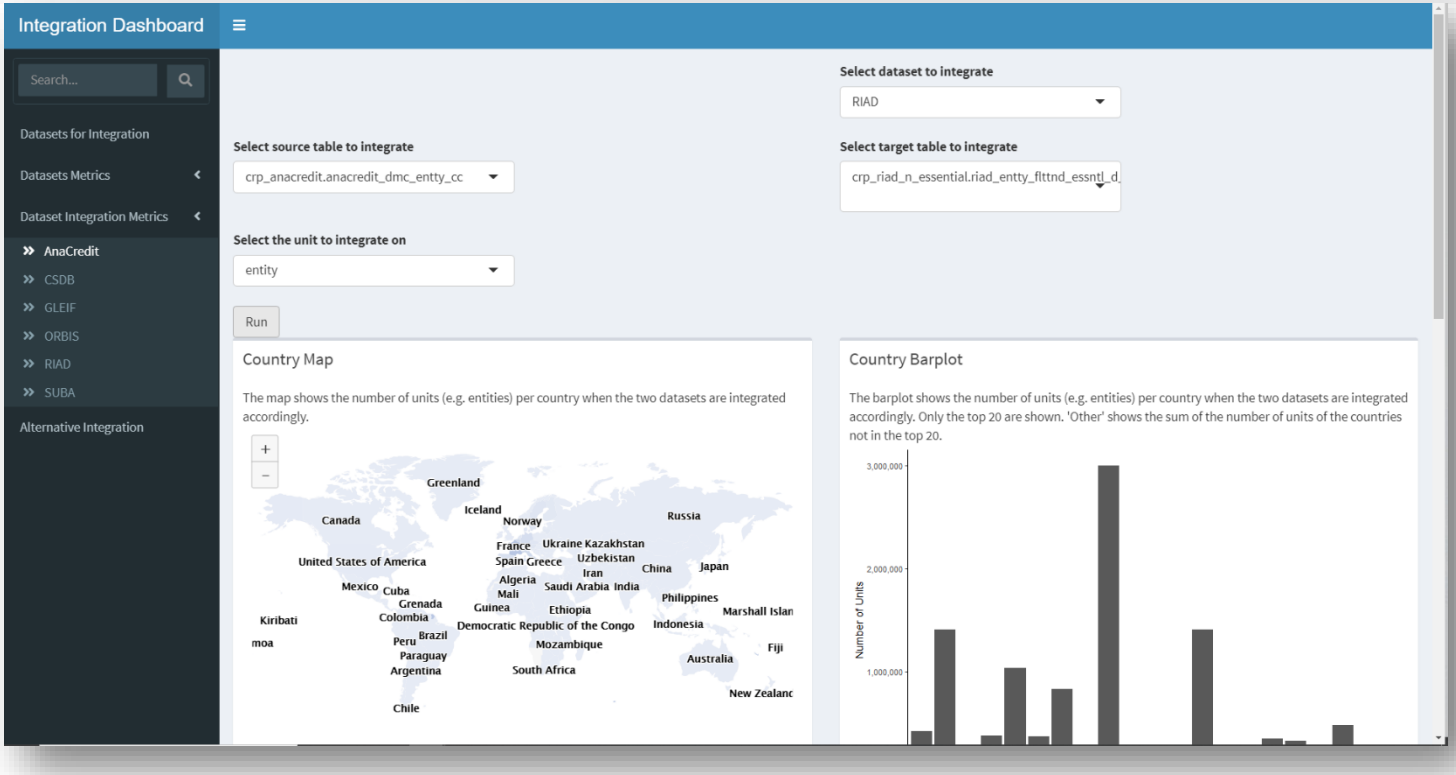
Note: due to confidentiality of some datasets, the visualisations are based on artificial data

Prototype dashboard – Details for a single dataset



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

Prototype dashboard – Integration of two datasets



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

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
Statistics Portugal

eurostat

The conference is partly financed by the European Union