

Reproducibility and (sensitive) granular data: What's the issue?

Stefan Bender, Head of Research Data and Service Center (RDSC), Deutsche Bundesbank

Love Your Code

Friday, 14 February 2020

UK Data Service and ONS

London, United Kingdom

Based on a joint project with and contributions from:

Stefan Bender, Jannick Blaschke, Hendrik Doll, Christian Hirsch,
Christian Resch, John Chase, Christian Herzog, Jonathan Morgan,
Ian Mulvaney, Andrew Gordon and Julia Lane

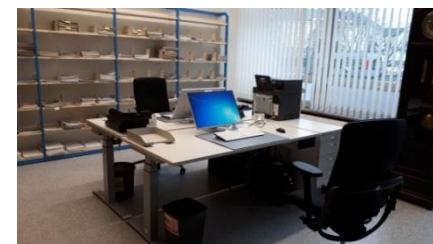
Tasks of the Research Data and Service Center (RDSC)

The RDSC offers access for non-commercial research to (highly sensitive) micro data of the Bundesbank. Microdata for banks, companies, securities and households are available:

- Generate (linked) micro data
- Offer advisory service on data selection and data access (data handling, research potential, scope and validity of data)
- Provide data access and data protection
- Document data and methodological aspects of the data
- Work on own research projects (in close cooperation with the Bank's business areas and the Research Centre)
- Organize conferences and workshops.

Factsheet on the RDSC

- 20 employees
- 12 working places for guest researchers in Frankfurt (fully booked several times).
- 2 working places in Düsseldorf
- In 2018:
 - Around 130 project applications, 73 were realized
 - Over 2,000 files (over 3.5 million lines) checked (output control)
 - Average of used data products per research project: 2.68
 - Papers of RDSC users are out
- In 2017 over 300 active projects, over 160 institutions involved (around 90 non-German).



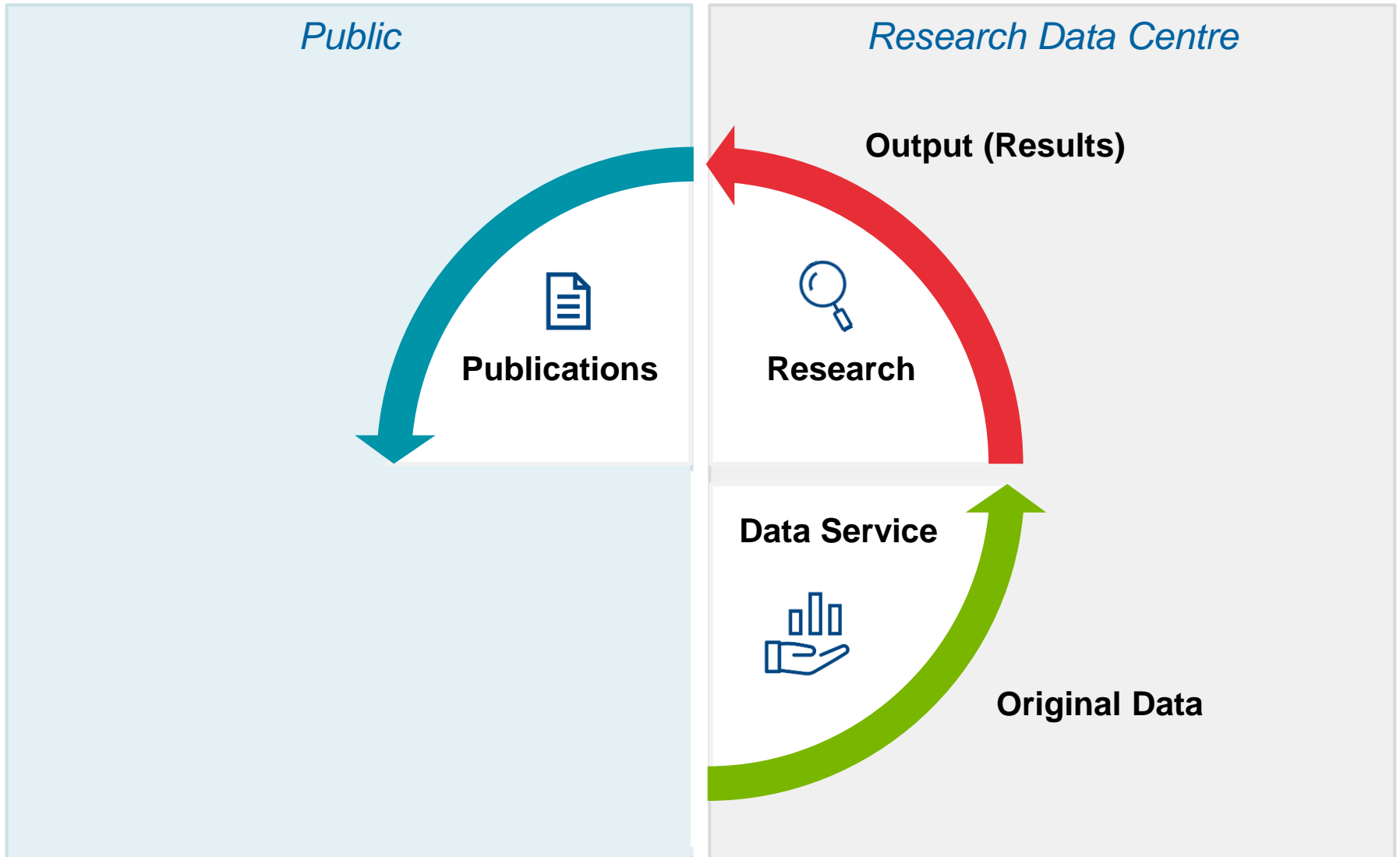
The Challenge - Part 1



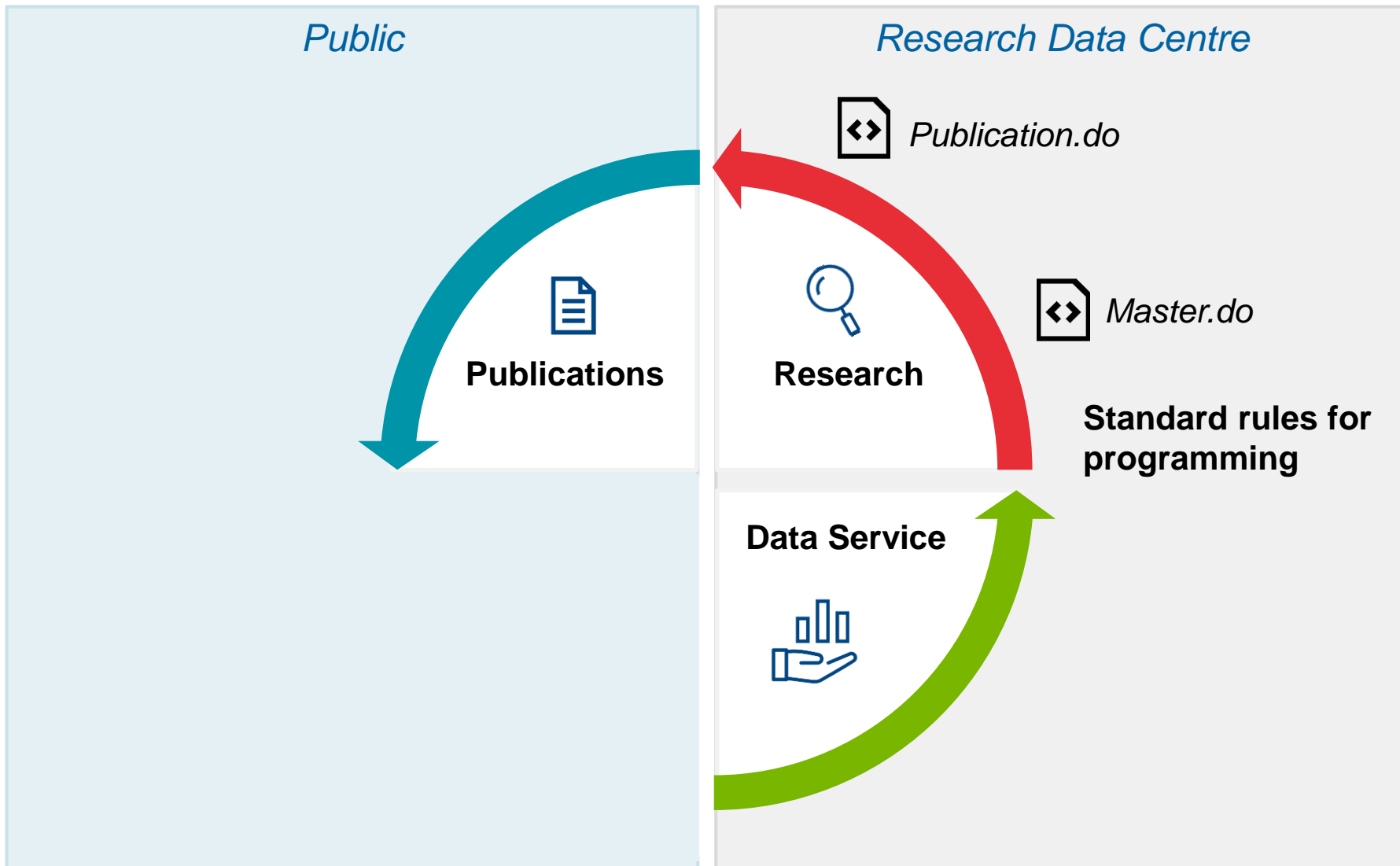
The Challenge - Part 1



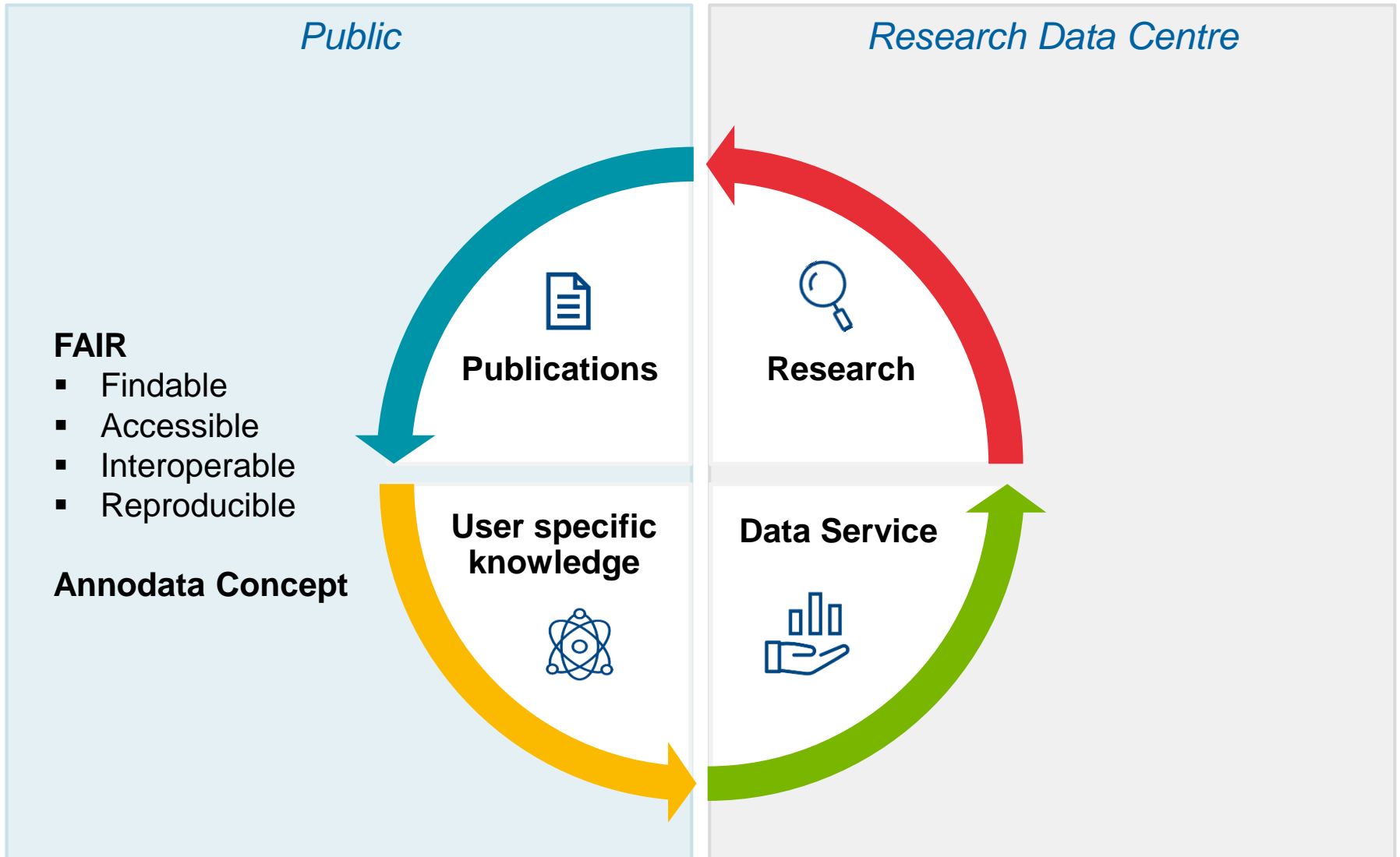
The Challenge - Part 1



Some Possible Solutions for Challenge - Part 1



The Challenge - Part 2 and Some Possible Solutions



Annodata-Schema

1



Access regime

2




Database

3



Dataset family

4




Record linkage

5



Combining restrictions

6



Global rules

7



Research projects

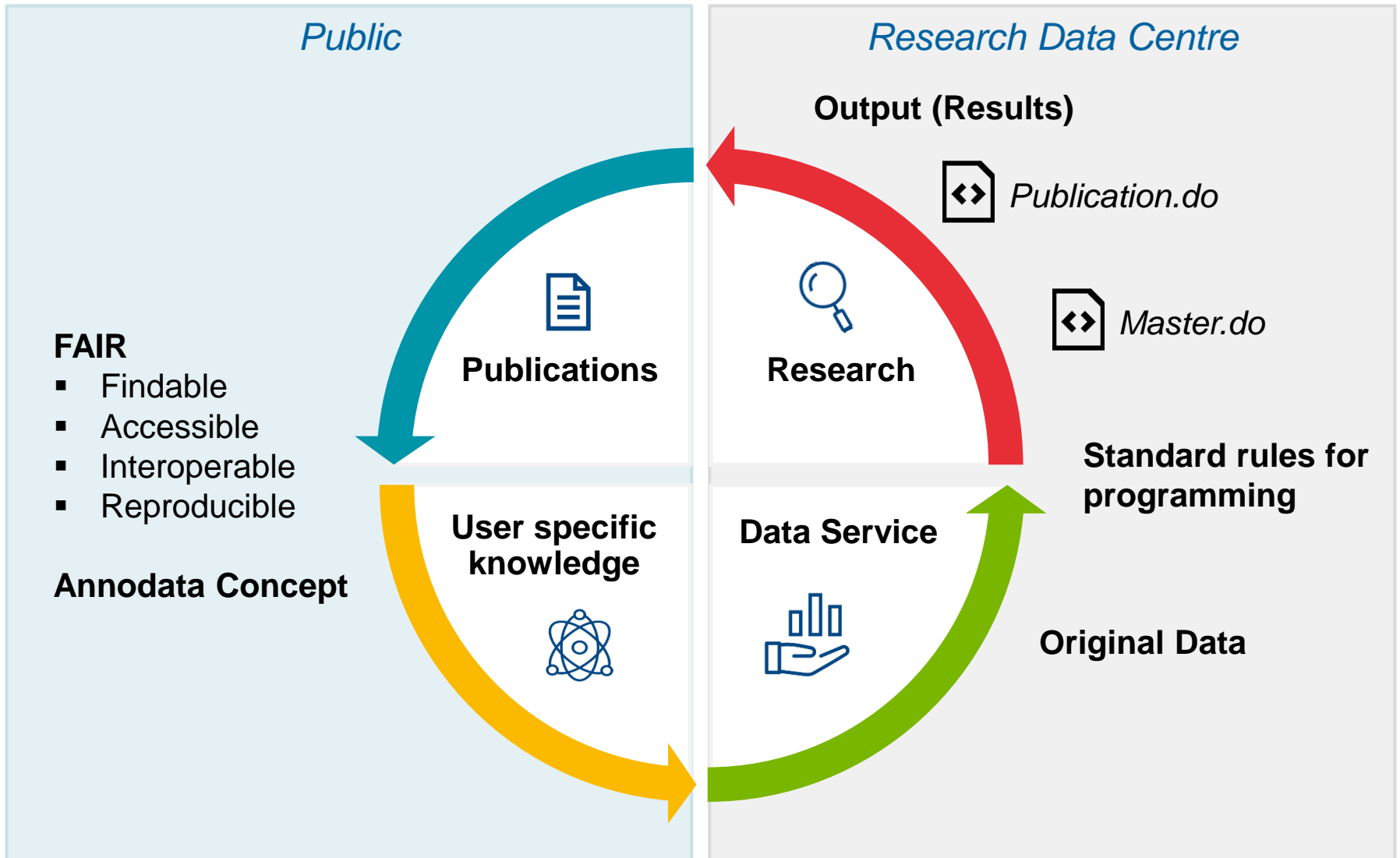
8



Researchers

- 1-3** on *dataset family* level
- 4-6** on *global* level
(i.e. irrespective of researcher affiliation, research field, and access mode)
- 7-8** on *project* level

Summing Up – but not the End



Data Sets in Publication: What has been done so far

1 Worldwide **competition*** with participants from all over the world and with New York University in the lead

- Meaningful solutions:

The logo for AI2, featuring the letters 'AI2' in a stylized blue font with a yellow dot above the '2'.

Allen AI



KAIST

The logo for gesis, with the word 'gesis' in a lowercase, blue, sans-serif font.

GESIS



Paderborn University

- ➔ First step for a systematic approach to find data sets in publications.

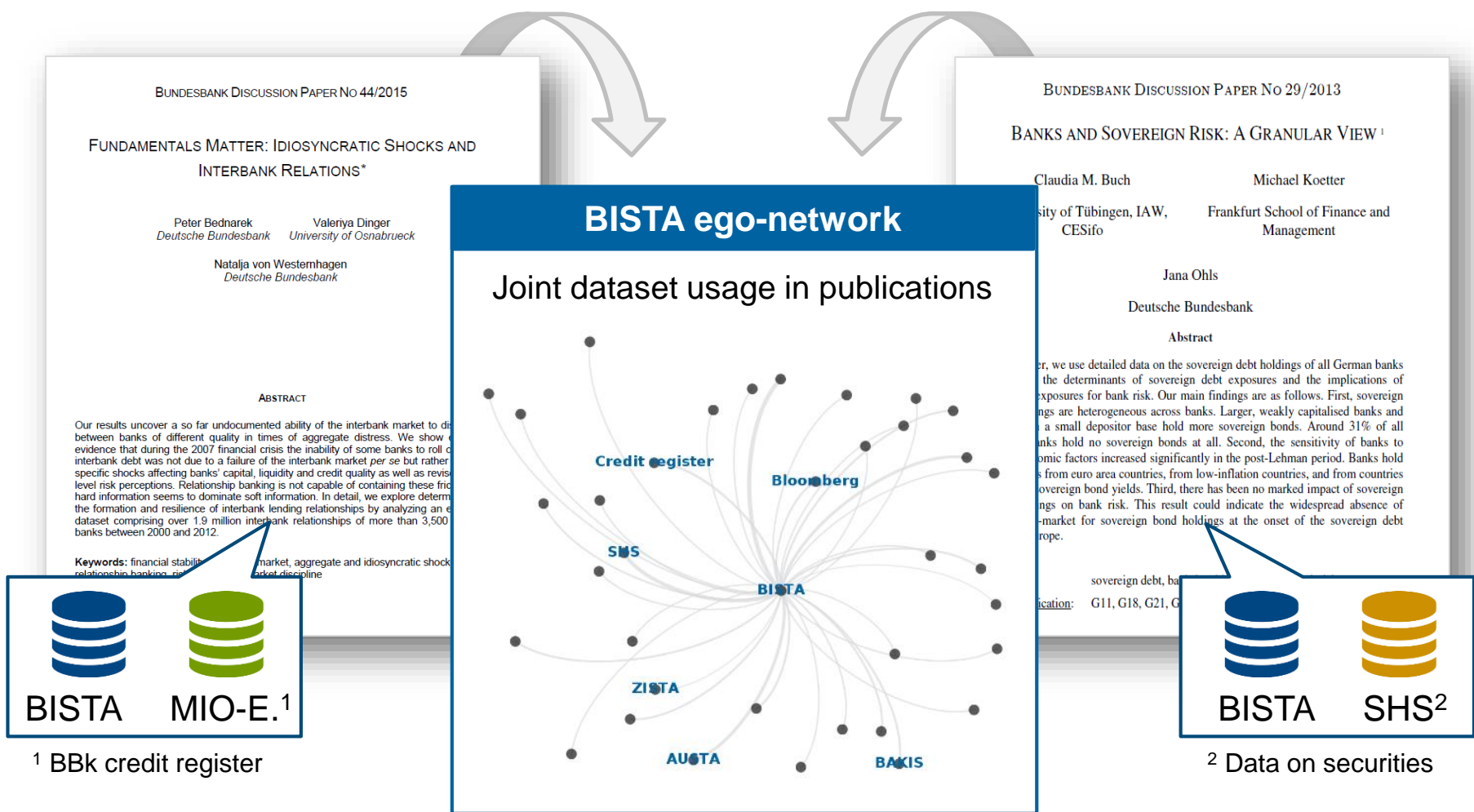
2 Contact with **RePEc** who think about including dataset mentions into their system

3 „**Rich Context Workshop**“ at National Press Club in Washington, DC to build a scientific basis for the empirical foundations of data science in government.

* For more information see <https://coleridgeinitiative.org/richcontextcompetition/workshopagenda>

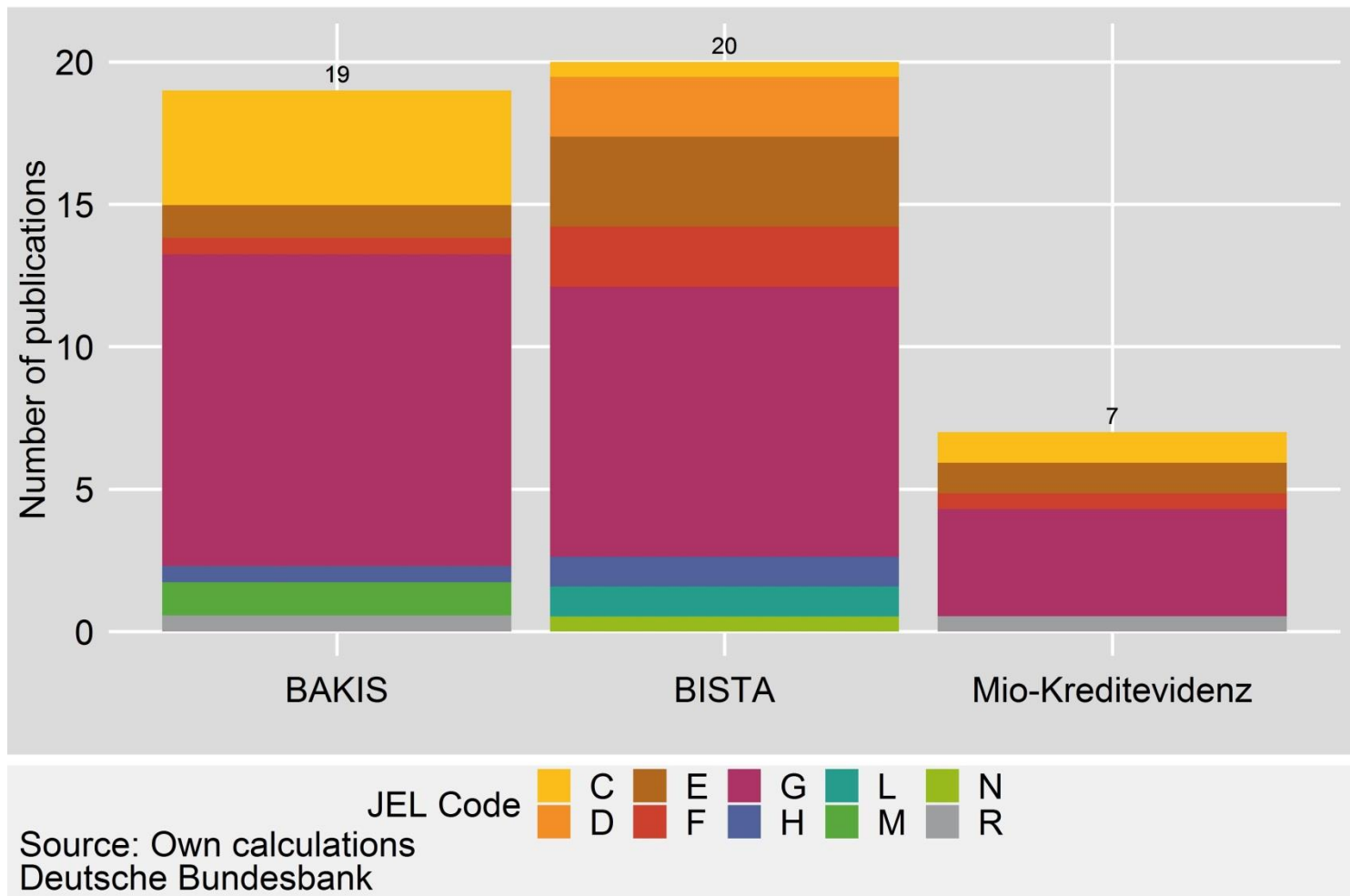
From “unrelated” articles to data network

Example: BBk’s monthly Balance Sheet Statistics (BISTA)



From dataset network to dataset impact factor

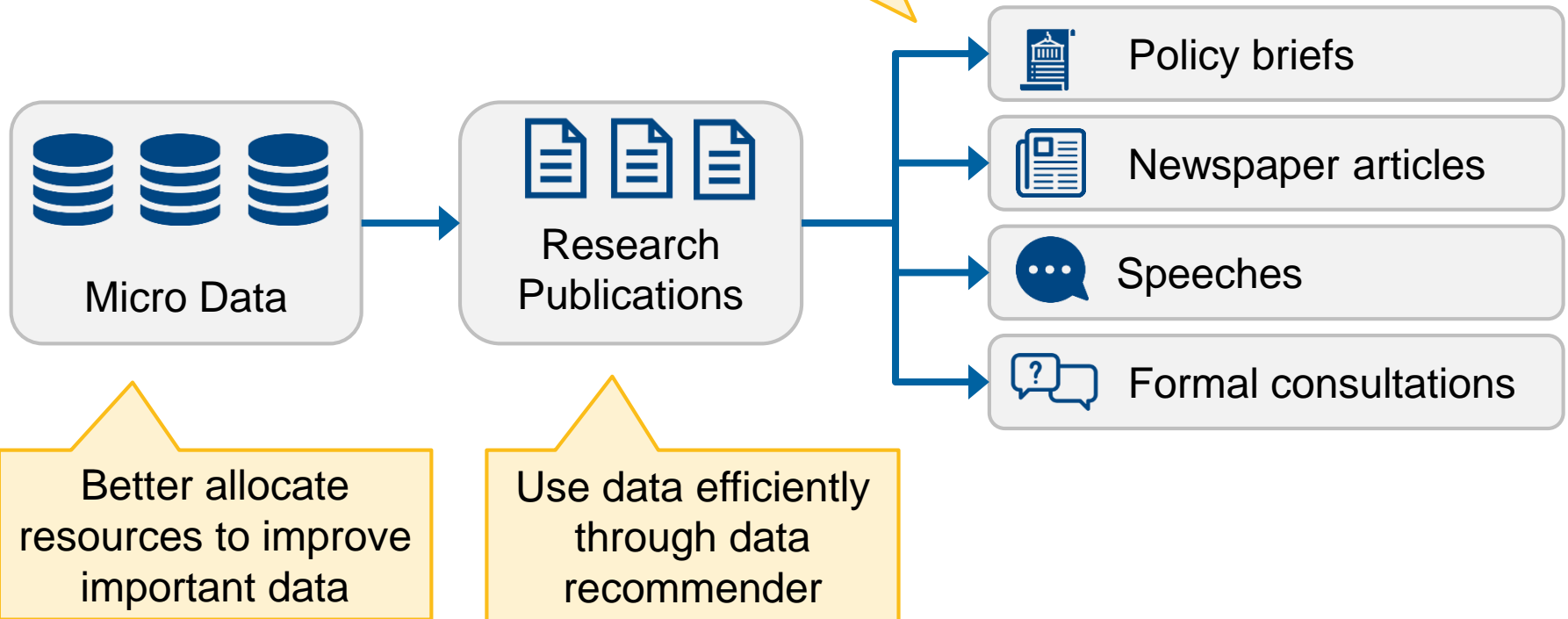
Number of publications per dataset and JEL code



Implications

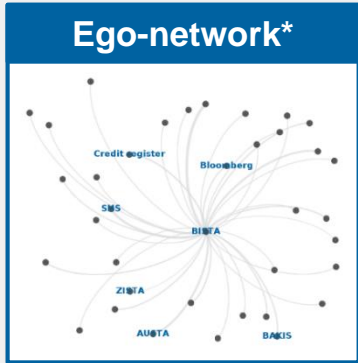
Evidence-based policy

Measure data impact on political decisions



Knowledge Life Cycle in RDSC (Bundesbank)

Measure data impact on political decisions.



***Example:** Joint dataset usage in publications for the BBK's monthly Balance Sheet Statistics (BISTA)

- Policy briefs
- Newspaper articles
- Speeches
- Formal consultations

Use data efficiently through data recommender (from ego-network).

Collaboration

- Knowledge sharing
- Metadata

Secure workspace

- Services and Tools

Publications

Research

User specific knowledge

Data Service

Better allocate resources to improve data quality and service.

Data Stewardship

- Approval
- Monitoring
- Reporting