The Role of the Bundesbank Microdata Production in Times of Big Data:
The need for Data access and Data Sharing

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

Open Science Conference 2018
13-14 March, 2018 Berlin, Germany

(The views expressed here do not necessarily reflect the opinion of the Deutsche Bundesbank or the Eurosystem.)
Bundesbank is publishing over 80,000 time series
- Time Series Data Base of the Bundesbank
- Macroeconomic time series (Real Time Data)
- European System of Central Banks

The number of time series increases per year (5-10% increase)

Bundesbank has trillions of micro data and around 600 millions of (virtual) time series.

One of the central bases for monetary policies and the analysis of financial stability.
- Aggregate datasets are important for monitoring macroeconomic developments and macroeconomic policy.

- However, they provide only an incomplete view of drivers and effects of changing structures in the real economy and financial sector.

- The analysis of heterogeneity is key to come to a better understanding of aggregated evolutions.

- Even more, granular data is necessary to understand global developments and in particular differences across countries.
Combining datasets and looking beyond aggregate statistics into heterogeneous developments require the transformation of “data” into “knowledge”.

Local constraints make it difficult, or often impossible, to link micro datasets from different jurisdictions, even for research and financial stability analysis.

Better accessibility and sharing of granular data would open up new possibilities for analysis by providing new insights into the effect of policies.

What can we do from the statistical side to support this process?
In 2009, finance ministers and central bank governors of the G20 endorsed the first Data Gaps Initiative (DGI-1) to promote actions to close data gaps related to the financial crisis.

G20 Data Gaps Initiative II (DGI-2), in particular recommendation 20, aiming to promote the exchange of data as well of metadata and addressing the accessibility of granular data.
Overview

- The Need for Giving Access to Granular/Micro Data
- Ethical Issues and Micro Data Access
  - Giving Access to Sensitive Micro Data
- INEXDA
- The German Data Forum
- “Conclusion”
Microdata at Bundesbank: The Need for Giving Access
Policy evaluation can make better use of existing datasets

- The Bundesbank – like other central banks – produces datasets which are highly valuable for policy analysis and research.
  - So far, most of these datasets have been used to provide aggregate statistics and ad hoc analysis of specific policy issues.
  - There is significant knowledge of data and institutional background.
- **Systematic use of these data for policy analysis is often constrained by**
  - Time
  - IT-resources
  - Legal restrictions
Motivation for establishing the RDSC: IMIDIAS

- The Bundesbank has launched a large-scale initiative aimed at making better use of existing data both, for policy analysis as well as internal and external researchers.

- The RDSC is part of the internal project Integrated MicroData-based Information and Analysis System (IMIDIAS)

Goals of IMIDIAS:

- Support policymaking process
- Encourage cooperation with (external) researchers
- Promote evidence-based policy-making
Available microdata at the RDSC

- Banks
- Companies
- Securities
- Households

Open Science Conference 2018, 13-14 March, Berlin, Germany
3/15/2018
Page 17
Available microdata at the RDSC

- **Banks**
  - Monthly balance sheet statistics (BISTA)
  - MFI interest rate statistics (MIR)
  - External position of banks (AUSTA)
  - Borrowers statistics (VJKRE)
  - Banks’ profit and loss accounts (GuV)
  - Banking Supervisory Data (available only with special restrictions)
  - Bank Lending Survey (BLS)

- **Companies**
  - Microdatabase foreign direct investments (MIDI)
  - Statistics on Trade in Services (SITS)
  - Corporate Balance Sheets (UStan)
  - Companies data from commercial data providers (if licence is available)

- **Securities**
  - Securities Holdings Statistics (WP-Invest)
  - Investment Funds Statistics (IFS)
  - Centralised Securities Database (CSDB; available only with special restrictions)

- **Households**
  - Panel on Household Finances (PHF)
  - Financial Literacy Survey (FLS)
Monthly balance sheet statistics (BISTA)

- The monthly balance sheet statistics list domestic banks (MFIs) assets and liabilities based on the books at the end of the month
  - Monthly information on loans and credits to other MFI, companies as well as households, containing credits, debt securities etc.
  - Unit of analysis: domestic MFI and foreign MFI acting in the domestic financial markets
  - Period and frequency of data collection: starting monthly in January 1999 until May 2017 as a panel
  - Number of units data is collected from: over 1,900 MFI
  - Number of variables collected: over 3,700 unique variables (with all positions according the German Commercial Code (HGB))
Microdatabase Direct Investment (MiDi)

- Information on inward foreign direct investments (FDI) as well as outward FDI
  - Granular information on FDI from domestic companies to companies located in other countries and incoming FDI from foreign owned companies to domestic and foreign owned companies
  - Statistical units: reports that contain the investment relationship between the transaction parties
  - Period and frequency of data collection: yearly reports starting in 1999 until 2015
  - Number of units data is collected from: over 440,000 annual reports
  - Number of variables collected: over 160 unique variables
  - Micro data is available as a panel
Bundesbank’s relevant microdata sources and their connections (excerpt)
Ethical Issues:
Access to Micro Data in a Big Data World
The Need for Data Access and Transparency in the Big Data Ages – The Role of Bundesbank

- **Data Access key principles:**
  - Data as a public good
  - Transparent data access
  - Data protection

- **CBs are in charge to give access** to “Big Data like” data for the wealth of society (also because of Big Data)

- **New form of transparency** about the results generated by CBs with data (reproducibility)

- But: keep **Trust**!
Types of Information considered as Personal Information

(QB2. Which of the following types of information and data that are related to you do you consider as personal?)

- Financial information (e.g. salary, bank details, credit record) - 75%
- Medical information (patient record, health information) - 74%
- Your national identity number/ identity card number/ passport number - 73%
- Your fingerprints - 64%
- Your home address - 57%
- Your mobile phone number - 53%
- Photos of you - 48%
- Your name - 46%
- Your work history - 30%
- Who your friends are - 30%
- Your tastes and opinions - 27%
- Your nationality - 26%
- Things you do (e.g. hobbies, sports, places you go) - 25%
- Websites you visit - 25%
Trust in Institutions to Protect Personal Information

QB25. Different authorities (government departments, local authorities, agencies) and private companies collect and store personal information. To what extent do you trust the following institutions to protect your personal information?

- Health and medical institutions: 78% trust, 20% do not trust, 2% don't know
- National public authorities (e.g. tax authorities, social security authorities): 70% trust, 28% do not trust, 2% don't know
- Banks and financial institutions: 62% trust, 35% do not trust, 3% don't know
- European institutions (European Commission, European Parliament, etc.): 55% trust, 35% do not trust, 10% don't know
- Shops and department stores: 39% trust, 57% do not trust, 4% don't know
- Phone companies, mobile phone companies and Internet Services Providers: 32% trust, 63% do not trust, 5% don't know
- Internet companies (Search Engines, Social Networking Sites, E-mail Services): 22% trust, 62% do not trust, 16% don't know

(Special Eurobarometer 359: Attitudes on Data Protection and, Electronic Identity in the European Union, Survey 11-12/2010)
Correlation between Trust in Banks and National Public Authorities by Country

(Special Eurobarometer 359: Attitudes on Data Protection and, Electronic Identity in the European Union, Survey 11-12/2010)
Who are we?
Statistics’ users!
What we want?! More data!
When we want?! Righ now!

THX to Filipa Lima
The RDSC offers access for **non-commercial** research to (highly sensitive) micro data of the Bundesbank:

- 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 (fully booked several times)
- Over 300 active projects
- In 2017:
  - Around 130 project applications, 73 were realized
  - Around 250 output controls
  - Average of used data products per research project: 2.15
  - First papers of RDSC users are out
Data Access at the RDSC

- RDSC mediates between data producers and external users.
- RDSC controls for compliance with data protection regulations.

Data Producers
- Survey studies
- Official statistics
- (External/Big Data)

Data users
Balancing usability and confidentiality is key

Analysis potential, data anonymisation and data access

- Original data
  - For statistical purposes
- De facto anonymised¹
  - On-site use,² scientific use file³
- Completely anonymised
  - Available for general use

¹ Data access in accordance with section 16 (6) of the Federal Statistics Act (Bundesstatistikgesetz). Microdata may be provided to academic institutions for the purposes of academic research if these data can only be traced to their source with a disproportionately large amount of time, costs and labour (de facto anonymisation).
² Use only within the Research Data and Service Centre. Results are subject to a mandatory disclosure control.
³ Scientific use files are anonymised in such a way that they may be used on the premises of the academic institution requesting the data.

Deutsche Bundesbank
## Modes of Data Access

<table>
<thead>
<tr>
<th>Off-Site Access</th>
<th>On-Site Access</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email, encrypted (Scientific Use File)</td>
<td>Remote Execution</td>
</tr>
<tr>
<td>Factually anonymous</td>
<td>Weakly anonymous (= confidential)</td>
</tr>
</tbody>
</table>

Output control
The 5 Safes in the RDSC (Portfolio Approach)

- **Safe people**: non-disclosure agreement, contract (with penalty up to 60,000 Euro, publishing the name, exclusion from access up to 2 years).

- **Safe projects**: non-commercial research, project description.

- **Safe environment**: working places without internet connection, (cell) phone, photo, printer and drive.

- **Safe data**: (weakly) anonymized data.

- **Safe results**: output control, papers/presentations are checked.

- **Access to real data**, anonymization is only one dimension, others have more effects on data protection.
International Network for Exchanging Experience on Statistical Handling of Granular Data (INEXDA)

Stefan Bender and Christian Hirsch, Deutsche Bundesbank
On 6th January 2017, have launched the International Network of Exchanging Experiences on Statistical Handling of Granular Data (INEXDA), an international cooperative project to declare their willingness to further strengthen their cooperation.

Since its foundation, the following institutions have joined INEXDA as a member:
The following institutions have participated as guests in the INEXDA meeting in Paris in January 2018:

- Bank for International Settlements
- Banco Central de Chile
- Banco de México
- Office of National Statistics UK
- Österreichische Nationalbank
- Türkiye Cumhuriyet Merkez Bankası

The following external experts have participated and provided expertise

- Julia Lane (NYU)
- Brigitte Hausstein (GESIS)
INEXDA’s MoU

- INEXDA is governed by an MoU, that every member has to sign.

- General mission is to promote data sharing and data access.

- Promoting the G20 Data Gaps Initiative II, in particular recommendation 20, addressing the accessibility of granular data. INEXDA is mentioned in a G20 paper.

- Sharing of granular data between INEXDA members not part of this MoU.
Detailed Work Programme for the first two Years

1. Perform a comprehensive inventory of data in all member institutions using a unified INEXDA metadata schema (THX to GESIS).
   - Agreement on unified metadata schema.
   - Setup of a platform to collect and exchange metadata among (and in the future possibly also beyond) INEXDA members.
   - First effort towards harmonising metadata across INEXDA member countries.

1. Perform a comprehensive inventory of existing data access procedures.
   - ECB pilot collection of information on access for researchers.
   - Expertise by Julia Lane (NYU).
   - Setup of working groups.

2. Dissemination of INEXDA results.
   - Set up of INEXDA webpage.
   - Planned future conference participations and workshops.
Administrative Data Research Facility (ADRF)

- The (ADRF) provides a secure platform to host confidential micro-data. It is developed at New York University (NYU) by Prof. Julia Lane ([http://www.julialane.org/](http://www.julialane.org/)) and team.

- ADRF combines the business workflow of a research data centre with potentially interesting new ideas how to enhance user experience and engage researchers to contributing information about data.

- ADRF comprises of the following 5 modules
  1. Documentation module
  2. Collaboration module
  3. Security module
  4. Stewardship module
  5. Training module
Related to data you've viewed

New data similar to data you've used

What others have done with similar data (recipes)

Recipes like yours

Thanks to Julia Lane
The German Data Forum

The views expressed here do not necessarily reflect the opinion of the Deutsche Bundesbank or the Eurosystem.
1.1 Data Access in Germany: Historical Development
Where do we come from?

- Access to official statistics was virtually impossible, or very costly (1990s)
- Bottom-up initiative of the research community and data producers (1998)
- The first research data centers were established (2001)
- The German Data Forum was established by the Federal Government (2004)
- 31 accredited Research Data Centers (2018)
1.2 German Data Forum: Key Facts

- Advisory council to the federal government
- 16 members: 8 data producers / 8 data users from research, own business office.
- Result of independent initiatives from within the scientific community (bottom up)
- Facilitating access to high-quality data
- Development of a research data infrastructure for the social, behavioral, and economic sciences
- Accreditation of 31 research data centers
1.3 Two Pillars of Activities

<table>
<thead>
<tr>
<th>Advising</th>
<th>Networking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scientific advisory of policy makers</td>
<td>National and international networking of infrastructures</td>
</tr>
<tr>
<td>Influencing relevant legislation</td>
<td>Development and improvement of the research infrastructure</td>
</tr>
<tr>
<td>Representing the interests and needs of the social, behavioral, and economic sciences</td>
<td>Accreditation and harmonization of research data centers</td>
</tr>
</tbody>
</table>
1.4 Current Agenda

- Further improvement of the research data infrastructure
- Recommendations on data access
- Digital support for survey data collection
- Archiving and secondary use of qualitative data
- Advising of legislators and policy-makers
- International networking
1.5 Accreditation of RDC

- Quality assurance, standardization
- Identify and communicate data needs
- Data Producers
  - Survey studies
  - Official statistics
- Research Data Center
- Data users
- Research ethics, data protection
- Mediation of different interests
"The mining of personal data can help increase welfare, lower search costs, and reduce economic inefficiencies; at the same time, it can be source of losses, economic inequalities, and power imbalances between those who hold the data and those whose data is controlled."

(Acquisti 2014, p. 98)
Conclusion II

- **Development** of the data infrastructure was/is fast, but **incremental**: trust building, growing data complexity, learning process …

- (New) **skills** for researchers / data producers.

- **Engagement** of researchers (value of data work?).

- **Efficiency**: researcher passport, metadata system (with elements of tripadvisor, amazon), project management in a RDC, …

- **Harmonization/Internationalization**: G20 initiative on data sharing and data access of central banks. INEXDA network.

- **German Data Forum as a role model for others** (communities, countries).
Thank you!

- **Website**: www.bundesbank.de/fdsz
- **Contact**: fdsz@bundesbank.de
Mobility

- Movement of people
- From home to work

assume all go by car

THX to Piet Daas, Statistics Netherlands

@Home 🏡
„Please let us drive. I just had one beer, I swear!“

Source: nelcartoons.de
1. Who wants change?

2. Who wants to change?

3. Who wants to lead the change?

Source: Kevin McKeown / Quebec Meme
What’s new in (central bank) statistics?

- Micro data overhaul the traditional value-added chain in central banking statistics.
  - Traditional central banking statistics are collected for a **specific purpose**.
  - Micro data are collected only once and can be used for **multiple purposes**:
    - The statistical reporting burden declines.
  - **Data protection** becomes more challenging.

- **Technological innovations** have revolutionized the infrastructure for collecting, storing, and using micro-data.
  - Advanced knowledge in storage and organization of large (integrated) micro-data.
  - Improved tools for analyzing and processing micro-data.
  - Cheaper storage technologies.
  - Standardization.
Data Generating Process

Until now (in many cases): ad hoc generation of data for research.

RDSC has started to/with:

- Establishing standardized data products.
- Implementing RDSC data quality procedures.
- Documentation of data.
- Harmonization of data.
- Register data to get data identifiers (DOIs).
Additional Aspects and Arguments for a RDSC

- **Trust** in researchers needed

- **Data quality** will increase

- **More results** on needed content and topics

- **Better knowledge** on data and content (recruitment)

- „**Branding“**, „**Marketing“**
Privacy Challenges in the Big Data world I: Behavioural Economics (Acquisti)

- Privacy valuations are significantly context dependent.
- Potential privacy paradox: people want privacy, but do not want to pay for it, and in fact are willing to disclose sensitive information for even small rewards.
- Consumers’ are not able to exhaustively consider the possible outcomes and risks of data disclosures (bounded rationality).
Privacy Challenges in the Big Data world II (Barocas, Nissenbaum)

- **Transparency Paradox**: Plain-language notices cannot provide information that people need to make decisions about complex contents in big data.

- **Informed consent** is believed to be an effective means of respecting individuals as autonomous decision makers with rights of self-determination.

- **The Tyranny of the Minority**: The willingness of a few individuals to disclose certain information implicates everyone else who happens to share the more easily observable traits that correlate with the revealed trait. The volunteered information of the few can unlock the same information about the many.

- **Inference**: A lot can be predicted about a person’s actions without knowing anything personal about them (especially in a big data context).
| Conclusion |

- INEXDA provides a platform for exchanging experiences on statistical handling of granular data for central banks, national statistical institutes and international organisations.

- Supports the G20 process, especially the Data Gaps Initiative 2 recommendation aiming to promote the exchange of (granular) data as well as metadata.

- So far, focus has been on taking stock which granular data is available in member institutions using a unified metadata schema.

- Focus is gradually shifting towards harmonising metadata and exchanging experiences about data access procedures.
Summing Up: New Challenges

- Define a research question (what are we measuring?):
  Do not fall in love with the Data. Love the questions it can answer.

- Think about what data are available (transactional versus aspirational) and the measurement error (how are we measuring it?):
  The size of the data reduces the estimation error, not its biases. Quality is what matters.

- Link datasets (what are we missing?)

- Statistical approaches (how can we draw inference?)

- Address Privacy and Confidentiality/Ethics (are we protecting human subjects?)

- Need for access and training

THX to Julia Lane and Roberto Rigobon