Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

Ninth IFC Conference on "Are post-crisis statistical initiatives completed?"

Basel, 30-31 August 2018

# Creating comprehensive data worlds using standardisation[1]

Stephan Müller,
Deutsche Bundesbank

[1] This paper was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

# Creating comprehensive data worlds using standardisation[1]

Stephan Müller[2]

Keywords: Micro data, growing data worlds, data gaps, data integration, standardisation, harmonisation, SDMX, House of Microdata

Ever since the global financial crisis, the importance of micro data has been on the rise. The amount of data being collected is constantly growing and becoming increasingly varied. In a parallel development, statistical authorities are faced with a rising number of data providers. Against this backdrop, this short paper aims to illustrate ways of enhancing the usability of growing data worlds.

## Growing data worlds

Since around 2008, the global financial crisis has led to a sharp shift in data needs and a surge in user groups. Many new statistical surveys were created to meet the huge demand for statistical information that emerged after the crisis. For example, the following questions had to be answered:

- How heavily affected are investors in Europe?
- Who holds which government bonds?
- What is the degree of risk concentration?
- How healthy are euro area banks?

Answering these questions increased the amount of available data enormously. Substantial technological progress meant that the data supply also expanded during this period. Examples of new technology include automatic recording of process data, social networks and search engines, as well as mobile phones, tablets and so on. And this went hand in hand with much larger computing power and new analysis techniques like machine learning. These developments led to an exploding data supply. And finally there was a paradigm shift from a macro to a micro perspective in terms of identifying certain heterogeneities. Facing with growing demand, growing supply and a shift from a macro to a micro perspective, it is therefore no exaggeration to say that the data universe is exploding. But can we really take advantage of all these newly available data?

[1] The paper is largely based on the book "Measuring the Data Universe" by Reinhold Stahl and Patricia Staab.

[2] Deutsche Bundesbank, Directorate General Statistics, stephan.mueller3@bundesbank.de

## The data universe lacks order

The status quo is that, despite our exploding data universe, there are still yawning data gaps. So a pressing question now is: do we drill where the oil actually is or where it is easy to drill? The data universe still lacks order. For example, there is still nothing like a unique identifier – an actual barcode for information. There are, for instance, MFI code lists or the Legal Entity Identifier – but these are not really global or universal.

The lack of order is evident in the data universes of almost all companies and has given rise to countless initiatives since a large part of companies' data are stored in data silos. Examples of widespread initiatives are projects relating to data integration, business intelligence, data warehouses or big data. Additionally, chief information officers are being appointed to bring order into their companies' data world. But the results are often proprietary solutions in the respective industry branches or countries.

There is also much higher volatility in terms of evaluation requirements. One can say that the focus is no longer on the classical statistical production of pre-defined indicators. Instead data analysis is, more and more, being implemented on demand with a lot of information needing to be available at all times – especially since evidence-based policymaking is now central to the regulatory agenda. All in all, a new style of data collection with hundreds of dimensions has emerged, meaning that there are more data, several structures, little order, high complexity and only a few experts available to analyse the existing data. And while the data may look similar, they are in fact very heterogeneous. Also, progressing automation can only help to a certain degree. Automation itself is a useful tool for data processing, but does not really help in terms of understanding, analysing and handling data. Especially when it comes to recognising the relationships between various sets of data, the experts need to share their knowledge and to cooperate with each other.

## Data integration

The current situation as described and the challenges involved raise the following question. What concrete measures can be taken to increase the usability of existing data? An important task is data integration. The process of data integration can be broken down into three steps. Each step can be technically supported and automated to a certain degree. Throughout the process, the degree of standardisation is increasing constantly and on an ongoing basis.

The process of data integration starts with heterogeneous data from various sources. The first step is logical centralisation, meaning that the data are stored physically or virtually in a common system. Common procedures can be used for administration, authorisation and access. This level of integration is what is meant when speaking of the data lake.

The second step is a uniform data modelling method with an order system, typically a uniform language using the same concepts and terms. An example is the use of SDMX as a standard designed to describe statistical data and metadata. Rule-

based and automatable treatment of the data thus becomes possible. At this stage, one typically speaks of a data warehouse.

The third step is semantic harmonisation. Here, the same concepts, methods and code-lists are used to classify the data. This makes it possible to link the data, i.e. to actually integrate content. As a consequence, a common dictionary can be used. This part is definitely the most difficult. At the end, the integrated data are ready for linking and simplification.

As a simple metaphor for the three steps of data integration, imagine a high-rack warehouse. Step 1 is to simply store all items in the same storage location. Step 2 is to put all items on racks so that nothing is left on the floor. Step 3 is to label the racks using a uniform system.

However, those wishing to introduce data integration face a couple of challenges from the various stakeholders involved. Every stakeholder has his own agenda with his own requirements. For example, the existing IT standards for data integration in the IT industry are either branch-specific silo solutions or high-level formal frameworks. In addition, within most companies, silo thinking is more pronounced than interdisciplinary thinking. Data users are not interested in data integration and the production process in itself, they are only looking for a specific result. Challenges from outside are often associated with privacy and data protection issues as well as with a lack of direct incentives.

## SDMX and the Bundesbank's central statistics infrastructure

To show how the Bundesbank deals with the topic of data integration, it is expedient to use the example of its statistical value chain. The Bundesbank receives data from its registered partners. The data are checked and aggregated in individual databases and IT systems for several primary statistics – in various data silos as it were. This represents the first step of data integration. Data from external organisations and commercial data sources are not modified.

The second step in data integration is to store macro and micro data in the Bundesbank's central statistics infrastructure. There, all data are stored in a common system and in SDMX format. The SDMX format allows all data to be edited using the same tools, and data sharing is also possible. As mentioned before, the most difficult part of data integration is step 3, namely semantic harmonisation. This would allow us to properly link all the available data, and the Bundesbank is currently working on making this possible.

At the end of the value chain are the different ways of using, disseminating and publishing the data. For example, there is an "access portal" for partners; this provides access to the available data not only to Bundesbank staff, but also to those of the Federal Ministry of Finance and the Federal Financial Supervisory Authority.

The use of SDMX as an organisational structure in the world of statistics has huge potential when it comes to data integration. The above-described statistical value chain leads to a certain data world. This world contains specific statistical data and metadata from different business areas. SDMX semantically translates the data into a uniform language. The SDMX keys can be considered as character strings that enable each time series of a dataset to be identified uniquely and read by a machine. After semantic translation, every time series of a dataset has a systematic

designation and can be organised in a data warehouse and used on a platform-independent basis. Once the process of data integration is complete, all the data in the different data warehouses of the various organisations could potentially be accessed using a common software product via a technical interface. Even linking the data from the various data warehouses would then be possible.

A good example of standardisation in times of an exploding data universe is the Bundesbank's House of Microdata. It is based on SDMX and the central statistics infrastructure and constitutes a central microdatabase able to hold all microdata with high potential for analytical purposes. Direct access to the House of Microdata is only granted to internal Bundesbank users, on a need-to-know basis and in compliance with confidentiality regulations. External researchers are obliged to use the services of the Research Data and Service Centre. The House of Microdata – like the Research Data and Service Centre – is part of the Integrated Microdata-based Information and Analysis System (IMIDIAS) and enables bank-wide data integration and a common information model. For each dataset to be integrated, a potential analysis is conducted. Once a dataset is completely integrated, it can be connected along the corresponding dimensions, which have been coded using the same code lists. This is actually the third step of data integration, and the Bundesbank is currently doing its best to semantically harmonise its formally SDMX-classified microdata.

## Outstanding issues in the context of standardisation

The analysis of the current situation may throw up to a couple of questions that might be relevant in the context of data standardisation and harmonisation in the future. At the moment, no globally consistent code lists are available yet. But is a truly global standard possible? Data standardisation and harmonisation are being promoted around the world but sometimes it seems everyone is tweaking the standard a little bit to meet their own needs. Meanwhile, truly universal and unique identifiers are urgently needed. Without them, we will always start over again. Another question might be whether the open source approach is appropriate for future efforts. In this context one could ask: why only share the data but not codes, knowledge, methods and programs? And do the global approach and the open source idea comply with confidentiality constraints and the legal framework? What good is a data warehouse if I am not allowed to share the data? Finally, are we investing enough time and effort into data literacy in order to sustainably increase the value of the existing data? Current discussions often only focus on technical aspects. But what about users' ability to really understand and master the data? An investment in data literacy would also improve trust in statistical statements.

## Reference

Stahl, R. and Staab, P. (2018). Measuring the Data Universe. 1st ed. Springer International Publishing.

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

Ninth IFC Conference on "Are post-crisis statistical initiatives completed?"

Basel, 30-31 August 2018

# Creating comprehensive data worlds using standardisation[1]

Stephan Müller,
Deutsche Bundesbank

[1] This presentation was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

DEUTSCHE BUNDESBANK
EUROSYSTEM

# Creating Comprehensive Data Worlds using Standardization

**Stephan Müller, Deutsche Bundesbank**

## Are post-crisis statistical initiatives completed?
## - The data universe is exploding -

### Demand for new statistical surveys

**Banking crisis**
How heavily affected are investors in Europe?

**Sovereign debt crisis**
Who holds which government bonds?

**Banking union**
What is the scope of risk concentration?

**Low-interest-rate environment**
How healthy are euro area banks?

### Exploding data supply

**Data amount is growing constantly and rapidly**

- Automatic recording of process data (sensors, Internet of Things)
- Social networks and search engines
- Mobile phones and tablets

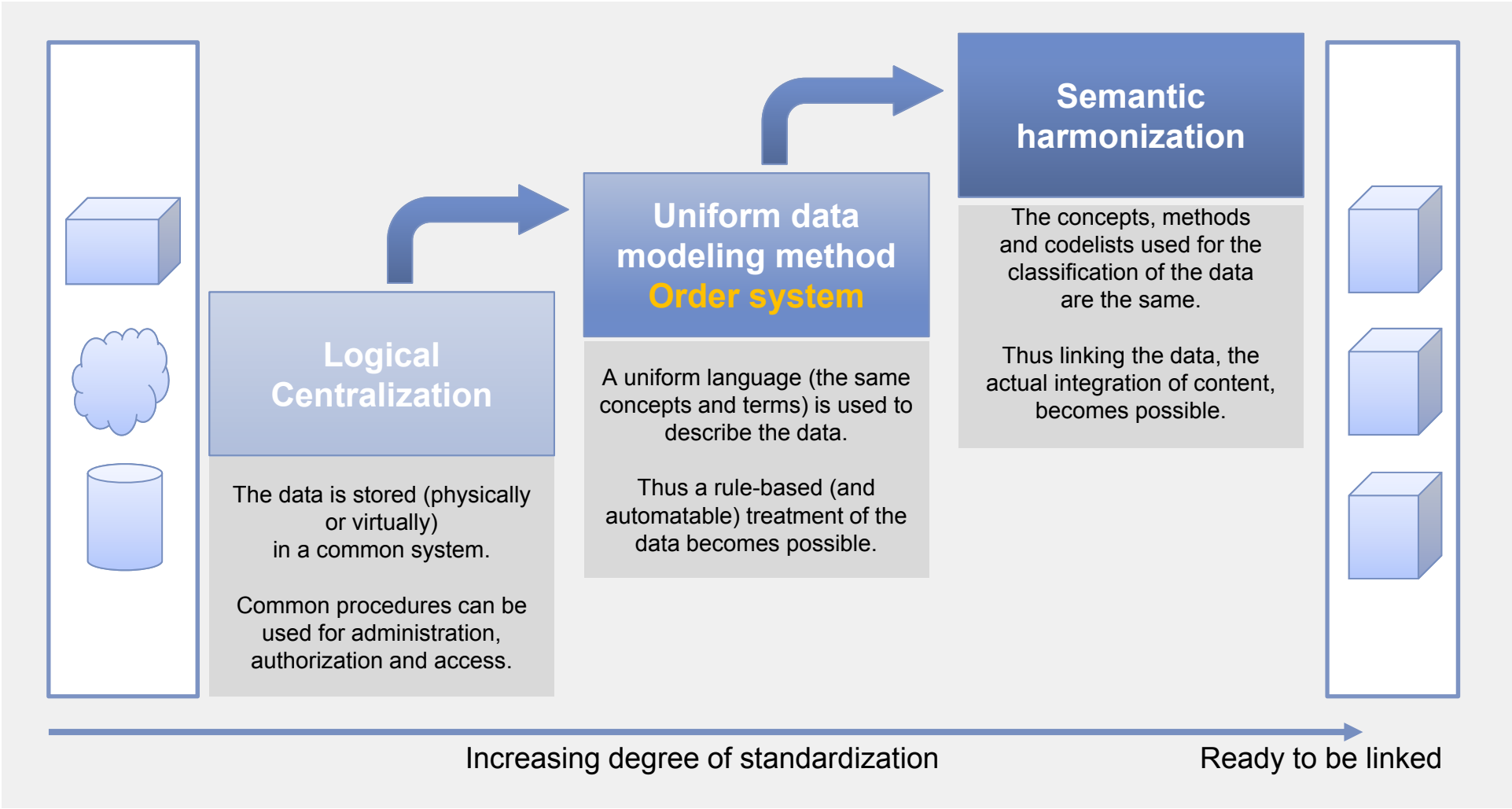
**New technological developments**

- More computing power: Big Data
- New analysis techniques: Machine Learning, AI

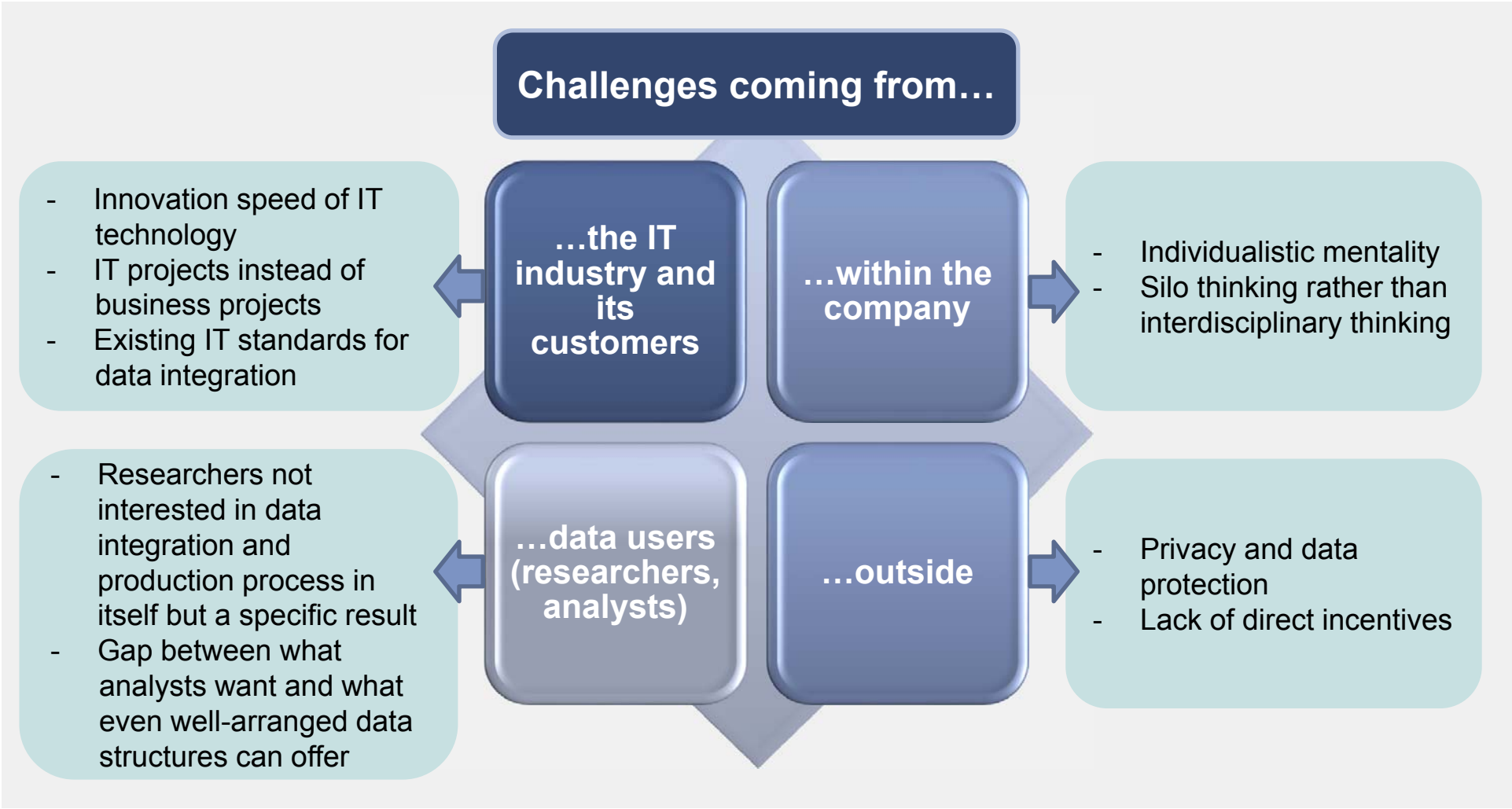
# “Water, water, everywhere, but not a drop to drink.”

- **Yawning Data Gaps despite “Collectomania”**
  - Data is not collected where it's needed, but where it occurs. Still painful data gaps
- **The Data Universe lacks Order**
  - In IT: neither a system of order for data / information, nor a prominent standardization, nor a global identifier (“barcode for information”)
  - In companies: large part of the data stored in data silos; need for data integration / BI / DWH / Big Data projects / CIOs
  - In industry branches or countries: proprietary solutions
- **Using IT not Possible Without Content-Related Expertise**
  - No longer classical statistical production of prescribed indicators
  - Instead implementation of data analysis on demand
  - New style of data collections with hundreds of dimensions
  - Automation or lack of expertise could lead to comparing apples and oranges
  - Professional expertise crucial for evaluating and interpreting the results

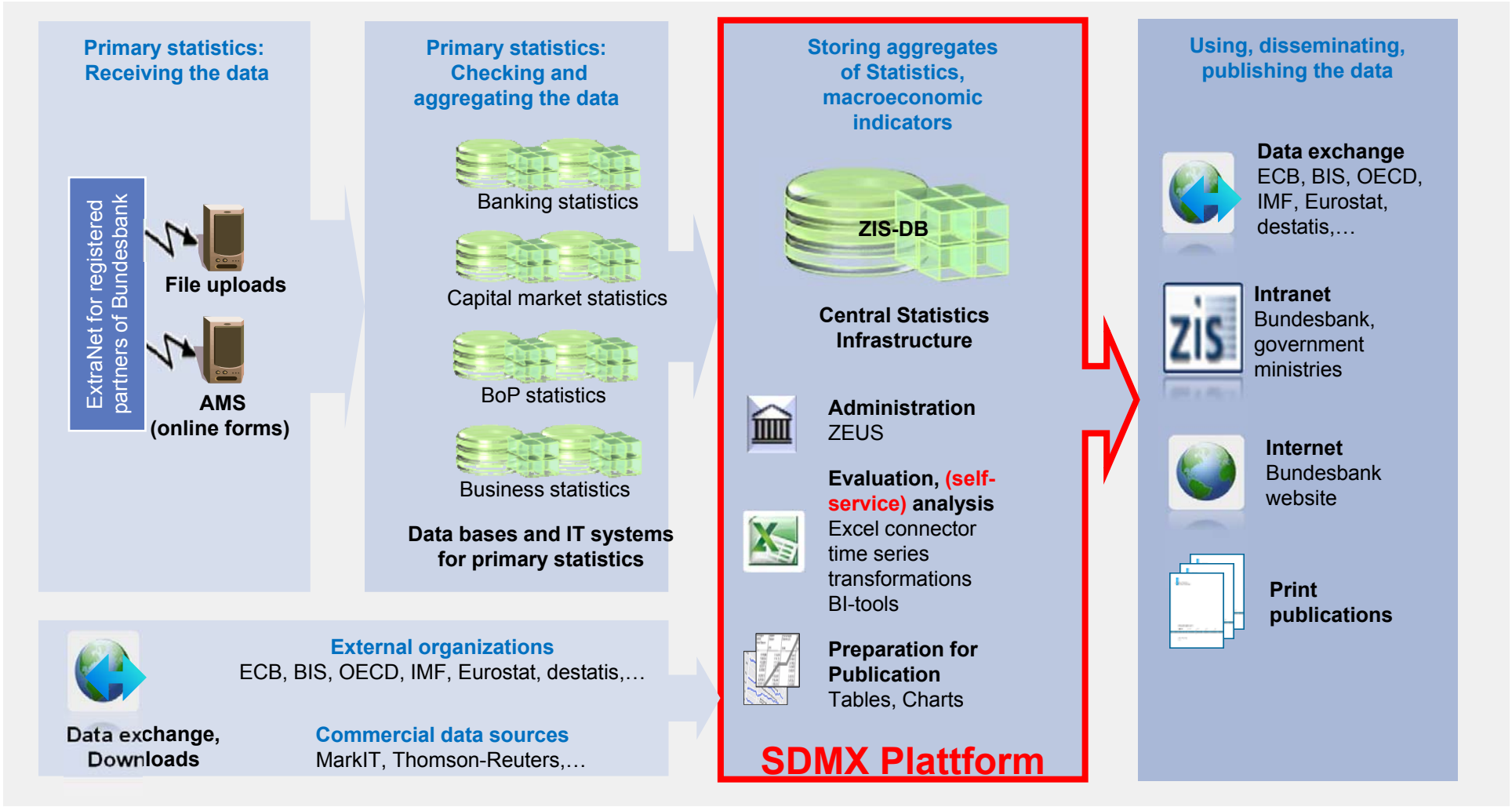
# The three steps of data integration



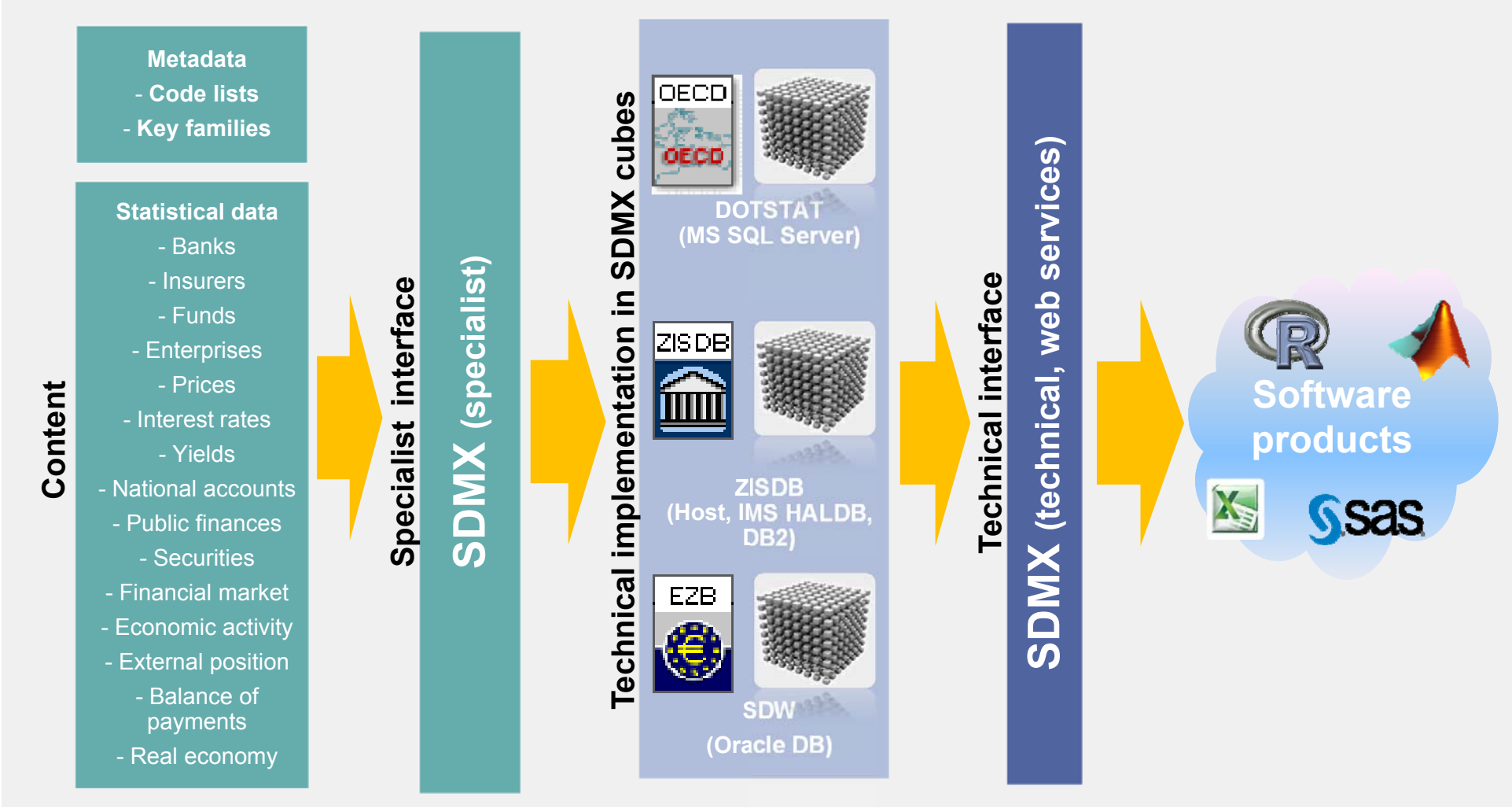
# Challenges for those who want to introduce data integration



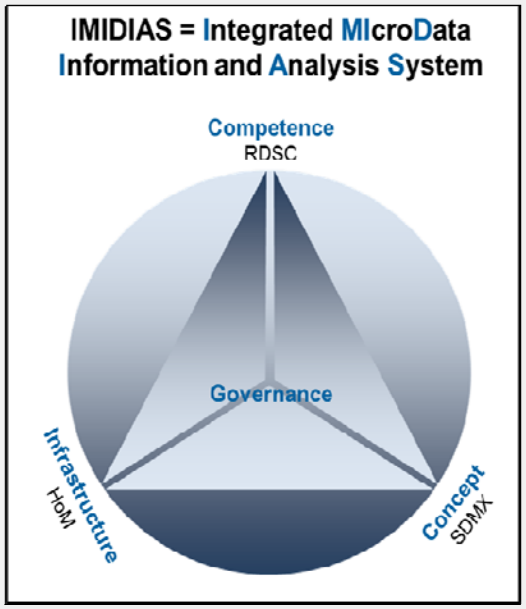
# Directorate General Statistics
# Value Chain



# SDMX is used across domains and platforms
# Decentralised data sinks on various technical platforms



# SDMX and central statistics infrastructure
# Basis for House of Microdata (HoM)




- In 2013, the Statistics Department was mandated to establish an **integrated interdepartmental information system for analytical and research purposes based on microdata** for various user groups (financial stability, research, monetary policy, supervision)
- This should be achieved by developing a Research Data and Service Centre (RDSC) and a **microdatabase (HoM, "House of Microdata")**
- This HoM is based on SDMX and the **Central Statistics Infrastructure**

− The SDMX model can be used without any problems for microdata.

− Data diversity requires standardization, SDMX provides a suitable framework

− Multidimensional approach, by using uniform code lists, offers an ideal means of linking and comparing data from different sources.

# What is there to do?

- There are **no globally consistent code lists** so far
  - Is a **truly global standard** possible?
- Is the global approach and the open source idea in accordance with **confidentiality constraints and the legal framework**?
- **Open source** approach for future efforts?
- Do we invest enough time and effort in **Data Literacy?**