The Bundesbank's House of Microdata

Standardisation as a success factor enabling data-sharing for analytical and research purposes

Patricia Staab
Deutsche Bundesbank, Central Office, Wilhelm-Epstein-Strasse 14, 60431 Frankfurt am Main, Germany

Abstract

To improve the sharing of microdata between Directorates General, the Bundesbank created a **House of Microdata (HoM)** on the basis of the existing statistical data warehouse infrastructure. The infrastructure uses the global statistics standard SDMX, which provides a powerful information *model for data-driven systems*.

By using coordinated code lists, the SDMX **generic multidimensional approach** offers an ideal means of linking, comparing and consolidating microdata. The HoM data can be used for any kind of dataset: for statistical as well as non-statistical data, and for aggregates as well as microdata.

Data diversity requires standardisation – the HoM provides a suitable framework.

The Bundesbank's IMIDIAS initiative

The Deutsche Bundesbank is one of the largest statistical data producers in Germany. It has collected monetary, financial, external sector and other statistical data for many years now. The data covering the fields of banks, securities, enterprises and household finance are used to produce **aggregated data** which are relevant for macroeconomic analysis. For decades, the microdata needed for this process were essentially viewed as just a by-product. Recently, this assessment has undergone a major change.

On the one hand, **research and analysis** focus more on understanding complex interactions between actors/sectors and market segments as well as the interaction between the financial sector and the real economy. Data providers are therefore faced with higher demands in terms of horizontal and vertical consistency, granularity and comparability.

On the other hand, several financial crises have seen **macroprudential issues** become increasingly important and a new core business of central banks. The questions arising from distribution and structural analyses as well as risk assessments require rapid and detailed examination of several datasets.

This demand cannot be met using aggregated data. In order to provide more direct access to the Bundesbank's **quality-tested administrative microdata** for

both policy analysts and researchers, the Bundesbank launched its large-scale **IMIDIAS (Integrated Microdata-based Information and Analysis System)** initiative.

In 2013, the Bundesbank's Directorate General Statistics received the mandate to coordinate the IMIDIAS initiative[1] and to provide its key components:

- The **House of Microdata**: A central data warehouse, based on a generic multidimensional data concept and able to hold all microdata with high potential for analytical purposes

- The **Research Data and Service Centre (RDSC)**: A service unit dedicated to supporting internal and external data analysts and researchers.

This paper focuses on the House of Microdata. For a description of the RDSC, see S Bender, and P Staab, (2015), as well as http://www.bundesbank.de/Navigation/EN/Bundesbank/Research/RDSC/rdsc.html.

The House of Microdata

As seen in Figure 1, all data stored in the HoM are extracted from pre-existing operating systems. The HoM stores **clean copies of the data**, thus leaving the pre-existing processes unchanged. Direct access to the HoM 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.
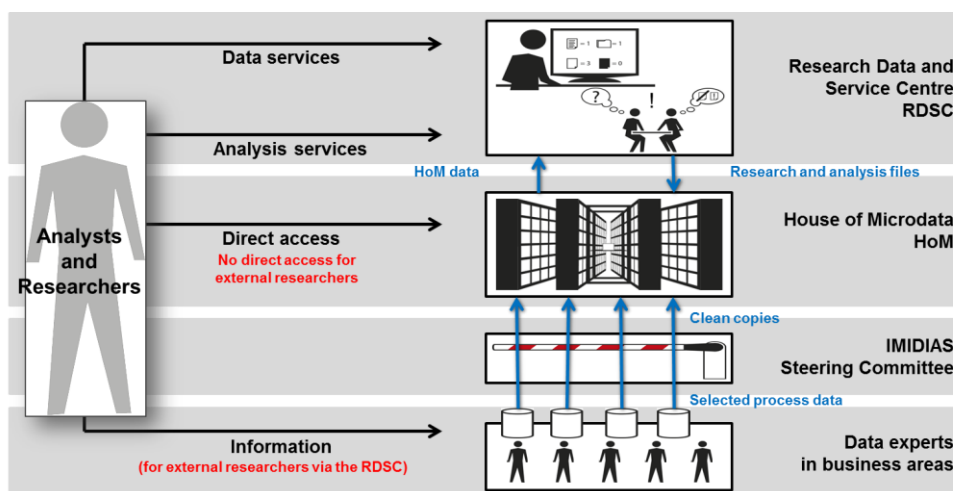
Figure 1: Schematic overview of IMIDIAS

Technically, the HoM is based on the **central statistics information system** already in use by the Bundesbank's DG Statistics. It consists of a data warehouse based on the international SDMX standard and its environment – several programs and interfaces to import, manage, search, present, analyse and export data.

---

[1] See also Irving Fisher Committee on Central Bank Statistics (2015), pp 40-41 for a description of IMIDIAS and the House of Microdata.

SDMX[2] is a global standard of the international statistics community, sponsored by its major institutions (the BIS, Eurostat, the ECB, the IMF, the OECD, the UN and the World Bank), and providing a data exchange format and generic multidimensional data model. The Bundesbank has been using this standard as a basis for its central statistics information system since 2003; a comprehensive statistical toolset for operational and analytical tasks is therefore already in place.

The central statistics information system has also been open to Bundesbank data pools outside the DG Statistics; several such datasets are already successfully integrated. It can largely be used independently of the DG Statistics and without the programming support of the Statistical Information Management Division. Standardised interfaces ensure that users can employ their own evaluation instruments rather than the information system's statistical toolset. In addition, a growing number of software products are available that make use of SDMX.

Statistical Standardisation via SDMX

From the 1990s up until the present day, several financial crises have led to the formation of global data initiatives under the guidance of institutions such as the IMF, the BIS, the FSB, the OECD, the ESRB and the G20. Besides the identification and elimination of data gaps, the initiatives also encouraged ongoing progress in data harmonisation: the creation of international statistical standards and codes as well as the streamlining of contents, methods, classification and procedures.

One such initiative was SDMX (Statistical Data and Metadata Exchange), which was launched by the sponsoring organisations the BIS, Eurostat, the ECB, the IMF, the OECD, the UN and the World Bank to standardise the international data exchange between statistical offices and central banks from 2001. SDMX did much more than just provide a universal data format for the exchange: it also created a **generic information model** for any kind of statistical data.

The core concept of SDMX is an astonishingly simple **building block approach**, in which a dataset is described as a multidimensional data cube. The number as well as the content of the cube's **dimensions** can be chosen according to the specific dataset. For each dimension, a list of possible entries – the **code list** – has to be defined. Then, any specific value within the dataset is identified by its **key**, which is the combination of corresponding dimension codes (see Figure 2). Dimensions, code lists and keys are called **metadata**, whereas the values are the actual **data**.

---

[2] For further information on SDMX, see the official site for the SDMX community at https://sdmx.org/
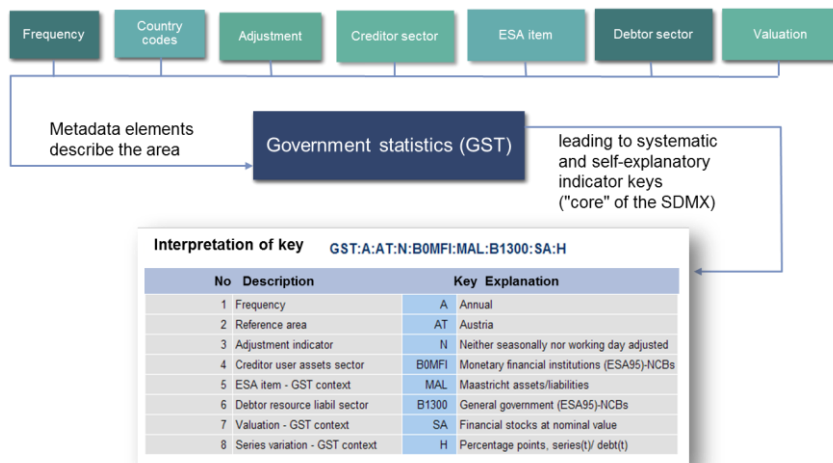
Figure 2: SDMX in a nutshell

The SDMX concept not only facilitated the creation of SDMX *self-explanatory* data structure definitions for all the datasets the sponsoring institutions needed to exchange. On top of that, it enabled the participants to implement *data-driven systems* to handle the data exchange. Establishing a new data flow between two institutions via SDMX consists of two steps (see Figure 3). In step 1, the Sender provides the metadata – the definition of the structure of the new data – to the Receiver. Once the data structure definition has been imported into the Receiver's SDMX-based system, the system is able to handle any SDMX dataset complying with the pre-defined data structure. In step 2, the Sender sends the SDMX data file, which is then processed accordingly.
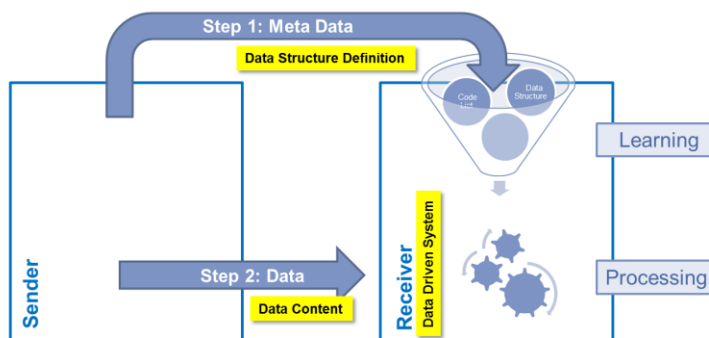

Figure 3: Data exchange between SDMX-based data-driven systems

Currently, SDMX (ISO standard as of 2005) is the standard used by the international statistics community not only for the **exchange** but also for the **publication of data**. The data exchange between the sponsoring organisations and their member states, especially the national statistics institutes (NSIs) and national central banks (NCBs), is based on SDMX. On top of that, many data portals as well as internal systems are organised as **SDMX data hubs**. The success of the standardisation initiative can therefore be seen on the websites of several international organisations. Just a few examples are the ECB Statistical Data Warehouse (http://sdw.ecb.europa.eu/), the OECD Statistics site (http://stats.oecd.org), and the Principal Global Indicators site (http://www.principalglobalindicators.org).

The SDMX building block approach makes it possible to store very different datasets side by side within the same data warehouse without having to integrate them semantically. However, integration is possible when datasets are defined in such a way that they **share a dimension**, ie they use the same definition of the dimension and agree on the same code list. Possible examples are the bank or security identifiers, but also the NACE classification codes.[3]

By sharing one or more dimensions, datasets are interconnected. An SDMX-based data warehouse can thus grow into a landscape of interconnected datasets (see Figure 4). The first building blocks of such a landscape have already been formed in the HoM.
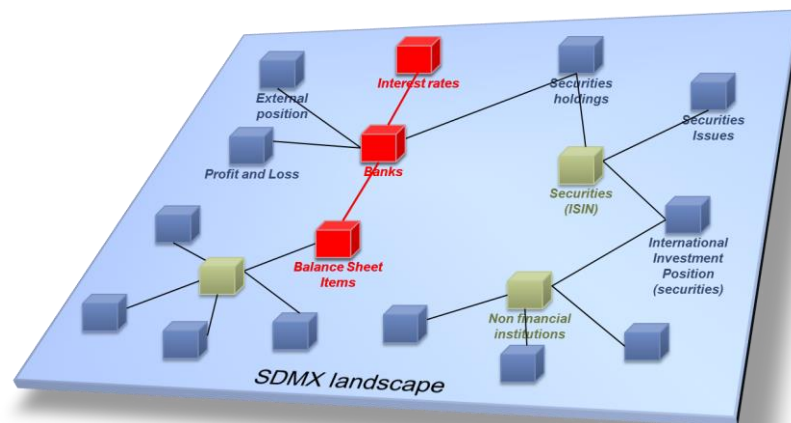

Figure 4: Example of an SDMX landscape

To conclude, the SDMX concept provides the framework necessary for data standardisation, while allowing for data variety.

- It is **generic:** The number and meaning of dimensions can be chosen freely; it can therefore be used for any kind of dataset (statistical and non-statistical data, aggregated and microdata).

- It is **multidimensional** by design and therefore offers the possibility of sharing dimensions and code lists. Thus, SDMX offers an ideal means of linking and comparing data from different sources.

A current example of an SDMX-based data warehouse is the Bundesbank's HoM. The successful integration of data into the HoM can be seen as proof of concept for the usage of SDMX. The following section describes the *status quo* regarding this data integration.

[3] International standard classification of all economic activities.

HoM: *Status quo* of data integration

All data to be integrated into the HoM are to be identified by the **IMIDIAS Steering Committee,** which consists of the data-using and data-providing DGs of the Bundesbank, as well as the DGs IT and Legal Services.

In 2014, the first coordinated and standardised **inventory list** of microdata sets at the Bundesbank, comprising the data of all relevant business units, was compiled under the Steering Committee's guidance. A **potential analysis** of datasets on the inventory list then led to the identification of **twelve data assets** to form the first wave of HoM contents. The main criteria used to assess the potential were relevance, sustainability, legal requirements and costs.

The IMIDIAS concept foresees a regular recurrence of the inventory list as well as the potential analysis; a second wave is to be launched in 2016.

For each data asset to be integrated, a content project was initiated. All content projects comply with a common framework (see Figure 5) and are coordinated by the IMIDIAS project office.
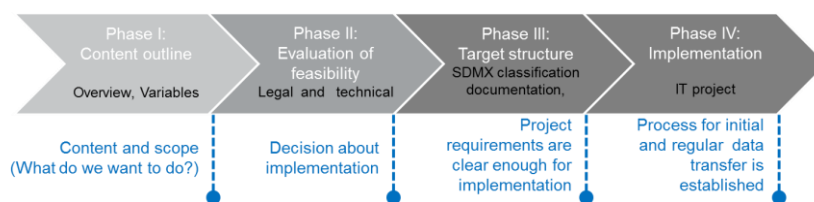
Figure 5: Phases of an IMIDIAS content project

Currently, nine content projects are at different phases of implementation and three content projects have been successfully completed.[4] The data sources already integrated are: Monthly Balance Sheet Statistics, Profit and Loss Accounts and Master Data of Banks. These data sources can be connected along the corresponding dimensions, which have been coded using the same code lists.

Conclusion

The **House of Microdata** is an integral part of the Bundesbank's **IMIDIAS** solution to address the current and future needs of researchers, analysts and policymakers. The underlying infrastructure is built upon the international statistics standard **SDMX** for data and metadata.

SDMX is **generic** and can therefore be used for statistical as well as non-statistical data, and for aggregates as well as microdata. SDMX is **multidimensional**; using coordinated code lists, it offers an ideal means of linking, comparing and consolidating microdata.

---

[4] To be completed: documentation of datasets for users.

The House of Microdata enables bank-wide data integration and a **common information model** - the key success factors are standardisation and harmonisation.

References

Becker, M, and A Treffs (2016). Establishing an integrated data sharing process for micro- and meta data at Deutsche Bundesbank. Retrieved from https://www.openconf.org/IASSIST16/modules/request.php?module=oc_program& action=summary.php&id=85

Bender, S and P Staab (2015). The Bundesbank's Research Data and Service Center (RDSC), Gateway to treasures of microdata on the German financial system. IFC Bulletin No 41, Irving Fisher Committee on Central Bank Statistics.

Deutsche Bundesbank (2015). Microdata – paradigm shift in central banks' statistics. Annual Report 2015, pages 47 to 59. Retrieved from http://www.bundesbank.de/Redaktion/EN/Downloads/Publications/Annual_Report/ 2015_annual_report.pdf?__blob=publicationFile

Irving Fisher Committee on Central Bank Statistics (2015). Data-sharing: issues and good practices, Report to BIS Governors prepared by the Task Force on Data Sharing, January 2015.

SDMX (2016). The official site for the SDMX community. Available at https://sdmx.org

Stahl, R (2016). Verlässliche Statistiken für das ESZB. Retrieved from https://www.bundesbank.de/Redaktion/DE/Downloads/Service/Meldewesen/Banke nstatistik/Anacredit/2016_06_29_praesentation_stahl.pdf?__blob=publicationFile

von Kalckreuth, U (2014), A Research Data and Service Centre (RDSC) at the Deutsche Bundesbank – a draft concept. IFC Bulletin No 37, Irving Fisher Committee on Central Bank Statistics.