Implementation of the SDMX International Standard in Statistical Data Specification

Maayan Kellerman and Elad David[1]

Abstract

Technological developments and the increase in the number of globally available data sources are leading to an immense increase in the volume of information and of the data inherent in it. In many cases, these data are stored and processed in various separate systems and databases, and are characterized by a lack of uniformity in data definition and how they are described – the metadata.

Similar to many entities around the world that deal with statistical data, the Bank of Israel is dealing with challenges in data decentralization and the lack of uniformity in their description by establishing a single database that brings together all data time series in the organization, creates uniformity in data definition, and standardizes the metadata, based on the SDMX[2] standard. The standard's purpose is to create uniform standardization of data description and to streamline data and metadata transfer processes and mechanisms between international organizations and between statistical organizations.

This paper discusses the importance of uniformity of metadata, presents the international SDMX standard, outlines the advantages of uniform modeling according to the international standard, and samples how the standard is implemented on data managed at the Bank of Israel.

[1] With thanks to the Information and Statistics Department's BI Unit for assistance in accessing the data on the dashboard for sampling purposes.

[2] Statistical Data and Metadata Exchange.

1. BACKGROUND

In recent years, due to technological developments and an increase in the number o available data sources, the world has seen immense growth in the volume, variety, and quality of data. As a result, the challenge of finding the necessary data and producing insights from them has increased.

In many cases, the data are stored and processed in various data systems and databases that are separate from each other (data silos), and feature a lack of uniformity in definitions and how they are described. In addition, it is difficult to link databases in order to obtain an overall integrative picture of all of an organization’s data. Meeting these challenges requires the creation of uniformity in data definitions and terms, and standardizing how they are described—the metadata—which is done through implementation of the SDMX standard. The standard was developed with the aim of creating uniform standardization for the description of statistical data, and streamlining the data and metadata transfer processes and mechanisms between international organizations, and between statistical organizations. The standard thereby helps implement data governance[3] in the organization.

The standard has so far been adopted by dozens of central banks, statistics bureaus, and international organizations. In recent years, the Bank of Israel has also begun to adopt it. The advancement of standardization of data description makes it possible to more precisely understand the data; enables extraction, filtering, sorting, integration, and linkage of data from separate systems, and makes it easier to query data and make international comparisons.

This paper discusses the importance of uniformity in the definition and description of data, presents the SDMX international standard, outlines the advantages of uniform modeling according to the international standard, and demonstrates how to implement the standard on the data managed by the Bank of Israel.

2. METADATA AND THE IMPORTANCE OF UNIFORMITY IN DATA DEFINITIONS

2.1 Description of the data – Metadata

Metadata is the information that describes the data. It is a combination of the prefix “meta” (Greek for “over”, “super”, “beyond”, or “after”) and the word “data.”

Metadata can be used to more easily identify, sort, and filter data that are managed in a fixed and defined structure. To illustrate, when you take a digital photograph, the picture that is produced is the data. The picture is accompanied by additional data that is gathered and saved: date and time, camera settings (photographic parameters), location, and more. These are the metadata of the photograph. The metadata makes it easier to sort pictures

[3] A collection of processes, functions, policies, and standards that aim to streamline the management and use of the data.

and to conduct searches and extraction of photographs from the database by metadata characteristics.

There are a number of common metadata characteristics in describing statistical data, including data frequency (whether the series is daily, weekly, monthly, quarterly, etc.), units of measure (percent, shekels, etc.). There are also other characteristics, some of which are unique to the particular content world.

Data without metadata, and metadata without data, lack significance. But the combination of data with their metadata enables both a better understanding of the data, and the ability to query and analyze the data by defined characteristics.

2.2 The importance of uniformity in data definitions and integration

The volume of collected data has increased markedly in recent years, due to the growth and availability of data sources, the increasing use of individual data, and the more frequent gathering of data.[4] If data gathered in the past were mostly aggregate and from a smaller number of sources, the recent financial crises have led central banks and data bureaus to understand the need to gather data at the individual or single-transaction level, and from a wider variety of sources and fields. The many technological improvements such as big data, data mining, and storage improvements, have also significantly improved data management, extraction, and analysis capabilities.

Gathered data are generally saved and processed in decentralized data silos that are built separately by each system manager on the basis of specifications of the unique needs of its users. While each database works efficiently on its own, and contains unique metadata, the individual definitions and separate terms and language in each database make it difficult for integration and linkage of the data in order to formulate a broader picture at the organizational level. For instance, definitions of activity segments may differ between databases. In one database, the definition of the "business sector" may include only real firms, while in another the term may also include financial firms. Since each database provides a different aspect of the information, integration between the data makes it possible to obtain an overall, fuller, and broader picture of the analyzed phenomenon, and helps achieve a deeper analysis of the data. In the above example, if the definition of "business sector" includes the same population of firms across databases, it would be possible to link the databases and carry out analyses that are based on a broader range of data, arriving at new and more in-depth insights.

[4] Data gathered in the past were at yearly, quarterly, or monthly frequencies, while data can now be collected on an intraday basis.

In addition to a lack of uniformity in definitions, working with decentralized databases may involve an overlap of data and a lack of consistency in the data that are presented. The more integrative the data management is, the greater the likelihood is that consistent data—economic phenomena and trends—will be presented in identical fashion, offering a single integrated message.

International experience[5] shows that in order to deal with the challenge of decentralized data and lack of uniformity, it is recommended to store statistical data in a single place, work toward creating uniformity in the definition of the data, and standardize metadata based on the international SDMX standard.

<u>Main characteristics of central vs decentralized database management</u>

Central database

Data providers

Data consumers

- All data series in one database;
- The system enables an overall and system-wide view;
- The data are modeled according to the SDMX standard to create uniform metadata;
- Users contact one database to extract, query, and link to data on the basis of a common language.

Decentralized database

Data providers

Data consumers

- Each database is built according to the users' specific needs;
- The database features itemized definitions in unique language and terminology;
- Lacks a holistic view at the organizational level;
- Difficult to integrate data;
- Users have to contact many information sources.

[5] For example: The Deutsche Bundesbank https://www.bis.org/ifc/events/isi_wsc_62/sts442_presentation2.pdf and the ECB https://www.bis.org/ifc/publ/ifcb49_36.pdf

3. THE INTERNATIONAL STATISTICAL DATA AND METADATA EXCHANGE (SDMX) STANDARD

SDMX (Statistical Data and Metadata Exchange) is an international initiative that aims to create a modern uniform standard for describing statistical data and streamlining the data and metadata transfer processes and mechanisms between international organizations and between statistical organizations. The initiative was developed in 2001 by seven major organizations involved in the collection and distribution of official statistical data, including the International Monetary Fund (IMF), the Bank for International Settlements (BIS), and the European Central Bank (ECB)[6] .

The initiative became an international standard adopted by the International Standards Organization (ISO)[7], and is now a globally accepted standard for exchanging data between organizations and for describing statistical data using uniform metadata.[8] The standard regulates how data is transferred (from machine to machine, machine to person, and person to machine), and thus streamlines and improves the quality of data transfer through standardization, automation, verification, and data sharing. In addition, the standard makes it possible to specify each data series through uniform metadata, as described below.

The SDMX is based on three main components: an information model that works to describe the data and metadata; statistical guides; and IT architecture and tools. The combination of these three components creates a common language between the business side and the IT side.

A. Information model[9] -

This model forms the core of the SDMX, and was developed in order to describe all of the data characteristics. The model combines the data themselves—the observations or measures—and the data characteristics and concepts, the metadata.

For instance, the figure 3.25, which was the shekel/dollar exchange rate on December 31, 2020, is only partially meaningful without its metadata. The significance of the figure is obtained from a description of the series' characteristics:

---

[6] The other sponsors were the Statistical Office of the European Union (EUROSTAT), the Organization for Economic Cooperation and Development (OECD), the United Nations (UN), and the World Bank.

[7] ISO 17369:2013

[8] Dozens of central banks and statistical bureaus around the world have adopted the standard.

[9] This section explains the main components of the model. There are additional components. For more details, see "Information Model: UML Conceptual Design".

Confidentiality classification	Multiple of units	Type of exchange rate	Type of figure	Counter currency	Base currency	Frequency
F – Unclassified	0 – Units	OF00 – Representative	E – End of period	USD	NIS	M – monthly

The SDMX model divides the characteristics into two classes: unique identifiers ("dimensions"), which make it possible to distinguish between series by combining dimensions; and other attributes—additional information on the data. In the above example, if we add data with different frequencies, currencies, figure types, and exchange rate types, the relevant categories would be defined as "dimensions", since they are unique identifiers. In other words, combining dimensions leads to one series. The unit multiples and confidentiality classification would be defined as "attributes", since they provide additional information on the series, but not for the purpose of identifying them.

There is a closed list of possible values for each of the characteristics in the example (code lists). For example, the code list for "counter currency" includes all currencies.

Combining the characteristics and the code lists of the values creates the structure of the content (data structure definition – DSD), where each statistical content world (financial statements from banks, real estate data from the real economy, etc.) has a separate DSD that contains data relevant to that content world. Each data array can be specified in the form of DSDs, where there are concepts that cross fields and those that are unique to a particular content world. The sponsors develop uniform lists of code lists, as well as global DSDs, so that all organizations specify their data in a uniform manner. For instance, balance of payments data have a global DSD called "BOP-Balance Of Payments", a frequency characteristic that includes a code list of data frequencies that is uniform, and so forth. International components are accessible and open for use by the broad public in a central place—the SDMX Global Registry. [10] The combination of the observations themselves with the characteristics creates datasets that are exchanged between organizations and made accessible on the statistical organizations' websites.

b. Statistical guides - A series of guides and recommendations that aim to encourage and create a common integrative language between SDMX appliers, thereby maximizing uniformity. The guides provide recommendations and tools that are based on terms that are common across fields and organizations, as well as international lists and codes with common classifications and structures. For instance, the Modelling Statistical Domains in SDMX guide[11] outlines the general principles for planning and creating SDMX components stage by stage. The guide offers tools and encourages discussion with the aim of setting the optimal number of characteristics for each content world.

[10] https://registry.sdmx.org/data/datastructure.html

[11] https://sdmx.org/wp-content/uploads/Modelling-statistical-domains-in-SDMX-v2-201806.docx

c. **Architecture and IT tools**– Many IT tools have been developed to support the adoption and implementation of SDMX. Most were based on open code so that they could form a convenient infrastructure for continued independent development in the adopting organizations. Among others, tools were developed to retrieve data together with the metadata, create SDMX-supporting files (with preference to xml standard templates), store the SDMX components, map and verify data structures, convert data between databases, and more. SDMX provides the necessary architecture for connecting IT systems to it, and thereby enables much easier data sharing.

Advantages to implementing SDMX in an organization:

1. **Improved coherence** - The use of metadata based on standardized terms, statistical guides, and the uniform structure for outlining content worlds based on the information model, make it possible to describe data more consistently and to better understand and interpret the data.

2. **Integration between data** - The use of uniform components for all content worlds makes it possible to link data and fields and obtain an over-all picture, which was not possible until the standard was adopted.

3. **International comparison** - The use of international values develops a common language across organizations and countries, and thereby increases the ability to compare with the data of other countries.

4. **Streamlining, resource conservation** - The use of identical data description structures while sending data in reporting files with a uniform structure and using open code software reduces development and maintenance costs.

5. **Accessibility** - The ability to rapidly extract and transfer data between organizations as a result of the adoption of the SDMX, which improves access and availability of the data, and enables rapid multidimensional querying.

6. **Reliability and reduction of errors** - Agreed-upon structures for the transfer and verification of the data and of automatic content reduce the need for manual intervention.

7. **Improvement of data exchange** - Data exchange is done in a common fixed data structure, with common templates, automatic publishing, and the ability to automatically extract information.

4. IMPLEMENTATION AND DEMONSTRATION AT THE BANK OF ISRAEL

Similar to many other statistical organizations around the world, the Bank of Israel has gained an understanding of the need to establish a single central platform for all statistical series managed at the Bank, and to create uniformity in the definition and modeling of the data according to the SDMX standard.[12] These processes are part of the data regime concept, and are intended to improve the value that the Bank derives from the data it manages, as well as their value for the data's users. This section outlines the stages of implementation at the Bank of Israel, and demonstrates some of the advantages in implementing the SDMX standard on the extraction, querying, linking, and comparison abilities of the data.

Implementation stages:

4.1 **Establishing a central statistical database -** The process of establishing a new central database that brings together all of the time data series in one place, included the establishment of infrastructure, development and purchase of technological tools, and the conversion in stages of the series from all decentralized databases that the Bank gathers and manages, and transferring them to the central database. The process included the mapping of the relevant series in order to make them accessible and convert them to the new database through IT processes.

4.2 **Creating uniformity in definitions** - The Information and Statistics Department created a uniform dictionary of terms for the activity segments and for the economic instruments in accordance with the international statistical methodology led by the International Monetary Fund (IMF).[13]

In general, the goal is to create uniformity in data definitions that is based on the international methodology. However, such uniformity cannot be achieved in every data system. For instance, defining the segments in the "debt aggregates" content world differs from the international methodology. In the data aggregates, due to constraints in the data sources, the "nonfinancial business sector" was defined to include nonfinancial business firms and some financial firms, while according to the international methodology, financial and nonfinancial firms are to be separated from each other. In such cases, the data are documented in the series metadata.

---

[12] The standard has two main aspects—replacing data using a uniform template (data exchange), and describing the data according to the information model. The Bank of Israel's adoption of the standard focuses at this stage on the data modeling portion, and the possibility of expanding the application will be examined later on.

[13] The IMF is responsible for publishing the guides for compilation of national statistics, while other organizations, including the OECD and central banks, participate in building the methodological framework for the compilation and for data reporting, and then act according to the IMF's guidelines.

4.3 **Adoption of the SDMX standard to describe data (metadata) –** As part of the adoption of the SDMX standard, the series in the central database undergo a modeling process in accordance with the standard's information model, using international components and according to the SDMX statistical guides, as much as possible. If the use of the international components is not possible, for instance, code lists or concepts that do not exist in the international methodology, the domestic components are used according to an orderly method, with a broad view of all information and data concepts.

So far, thousands of modeled series have been moved to the central database, and in the coming years, the process for all statistical series existing at the Bank of Israel is expected to be completed. Extraction of the series is through a dedicated interface that will be accessible on the Bank of Israel website, and the series will also be accessible through a BI tool, statistical software, and API.[14]

Demonstration:

The main feature for identifying a series used to be the "series name". Use of the "series name" as the main identifying feature made it difficult to distinguish between various series, and limited sorting, extraction and querying capabilities. In contrast, in modeling the information according to the SDMX, the content structure (DSD) is defined to include a wide variety of features to describe the data. Some of the features are generic (such as frequency), and some are unique to particular content worlds. The following table presents a sample of modeling according to the SDMX of a number of series from the "debt aggregates" content world, which includes the main data description features.

Data type	Seasonal adjustment	Unit multiplier	Measurement unit	Borrowing sector	Lending sector	Instrument type	Frequency	Series name
Closing balance	Original data	Billion	NIS	Business sector excluding banks, insurance, and credit card companies	Institutional investors	Loans	Quarterly	**Nonfinancial business sector debt to institutional investors, loans**

[14] Application Programming Interface - which allows applications from outside entities to access the data and functionalities that the organization provides, in a simple and secure fashion.

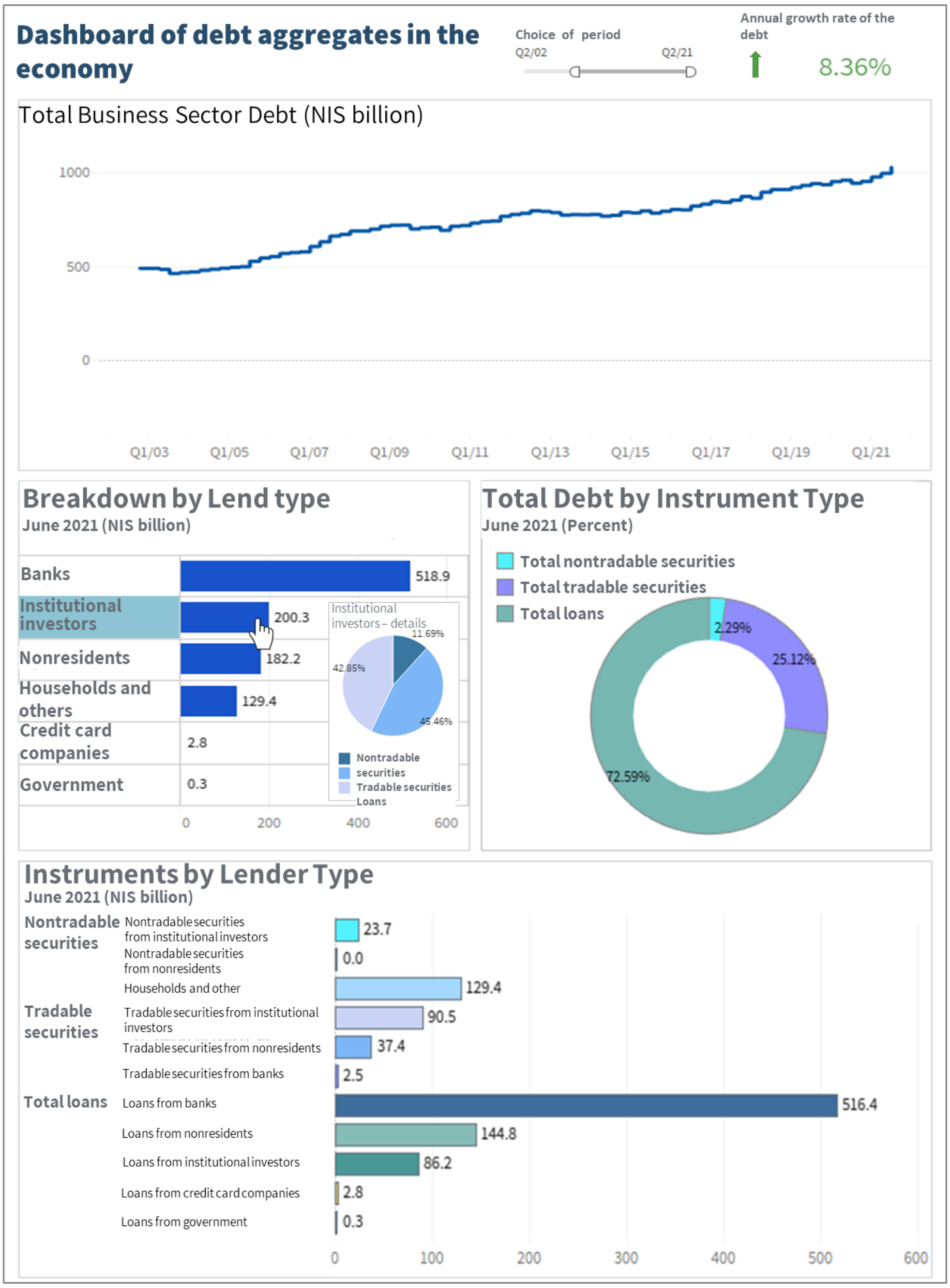
Closing balance	Original data	Billion	NIS	Business sector excluding banks, insurance, and credit card companies	Institutional investors	Nontradable securities	Quarterly	**Nonfinancial business sector debt to institutional investors, nontradable securities**
Closing balance	Original data	Billion	NIS	Business sector excluding banks, insurance, and credit card companies	Institutional investors	Tradable securities	Quarterly	**Nonfinancial business sector debt to institutional investors, tradable securities**

Information modeling through measurements enables a wider variety of breakdowns and makes it possible to analyze the data from a greater number of angles than with the use of "series name" as the main identifying feature. For example: breakdown by type of instrument or sector.

Another advantage of establishing a central database and multidimensional modeling is the possibility of using business intelligence (BI) tools that provide the ability to make data accessible visually, dynamically, and flexibly, and to conduct advanced queries using dashboards based on the features that were modeled.

Thus, the end user can make comparative surveys and trend analyses, change variables, and examine various phenomena easily and quickly. The following is a visual example of data specification according to the information model using a dashboard that presents the data by breakdowns based on the features that were defined in the information modeling stage. The dashboard includes a presentation of total business sector debt over time, and breakdowns by measures that were defined in the information model—instrument type and lender type. The dashboard is dynamic, and makes it possible to choose date ranges, change the focus by choosing features that were defined in the information modeling stage, which enables a dynamic analysis according to need.[15]

[15] The dashboard presented in this paper is for illustrative purposes only.



Dashboard of debt aggregates in the economy
Choice of period
Q2/02
Q2/21
Annual growth rate of the debt
8.36%
Total Business Sector Debt (NIS billion)
1000
500
0
Q1/03
Q1/05
Q1/07
Q1/09
Q1/11
Q1/13
Q1/15
Q1/17
Q1/19
Q1/21
Breakdown by Lend type
June 2021 (NIS billion)
Banks
518.9
Institutional investors
200.3
Nonresidents
182.2
Households and others
129.4
Credit card companies
2.8
Government
0.3
0
200
400
600
Institutional investors – details
11.69%
42.85%
45.46%
Nontradable securities
Tradable securities
Loans
Total Debt by Instrument Type
June 2021 (Percent)
Total nontradable securities
Total tradable securities
Total loans
2.29%
25.12%
72.59%
Instruments by Lender Type
June 2021 (NIS billion)
Nontradable securities
Nontradable securities from institutional investors
23.7
Nontradable securities from nonresidents
0.0
Households and other
129.4
Tradable securities
Tradable securities from institutional investors
90.5
Tradable securities from nonresidents
37.4
Tradable securities from banks
2.5
Total loans
Loans from banks
516.4
Loans from nonresidents
144.8
Loans from institutional investors
86.2
Loans from credit card companies
2.8
Loans from government
0.3
0
100
200
300
400
500
600

5. Conclusion and future plans

With the increase in the number of information sources and technological developments, the volume of data managed by statistical organizations, including the Bank of Israel, has increased markedly. The data are in many cases characterized by a lack of uniformity in definitions and in how the data are described—the metadata. This paper explained what metadata is, discussed the importance of creating uniformity, and described how the Bank of Israel is moving from decentralized and nonuniform data management to managing data in a central database based on uniform definitions and data modeling according to the SDMX international standard.

These measures provide the ability to better understand and analyze the data, based on extraction, filtering, and sorting capabilities according to measures; the ability to integrate the data, based on which a broader range of data can be analyzed and new and deeper insights can be obtained; compare domestic data to data from other countries that implement the standard; access data using statistical software, API, and BI tools, which make it possible to access data visually, dynamically, and flexibly; and the ability to perform advanced queries using dashboards.

Future plans:

1. Complete the process of putting all the aggregate statistical series that the Bank manages into the series database, so that all content worlds will be modeled according to the SDMX standard and access to data will be improved through dashboards and a new series retrieval interface.
2. Examine the expanded application of the SDMX standard in the following areas:
   a. Modeling data from databases that include granular data – The need and the ability to manage itemized databases that include data on an individual or firm level have increased in recent years. Granular data modeling according to the standard will contribute to uniformity and integration between a broader variety of data.

   b. Developing the ability to exchange data and metadata with other organizations through the application of the standard in this field (data exchange).

6. References

1. Bholat, D. (2016). “Modelling Metadata in Central Banks”, Statistics Paper Series No. 13, European Central Bank.
2. Stahl, R. and P. Staab (2018). Measuring the Data Universe: Data Integration Using Statistical Data and Metadata Exchange, Springer International Publishing AG.
3. https://sdmx.org/
4. https://ec.europa.eu/eurostat/
5. https://en.wikipedia.org/wiki/Metadata