Design and Implementation of Scalable Data Collection Framework for SDMX

Taehwan Kim, Hyunjoong Kang, Taewan You, Yeonhee Lee
Fiscal and Economic Intelligence Research Center
Electronics and Telecommunications Research Institute
Daejeon, Korea
{thkimetri, kanghj, twyou, leeyh}@etri.re.kr

***Abstract*— With the increasing adoption of the SDMX (Statistical Data and Metadata Exchange) standard by major national statistical institutes and public authorities, researchers and general users now have greater access to high-quality and reliable statistical data. However, collecting and managing data from various SDMX open data sources presents significant challenges. Firstly, despite operating under the same version of the SDMX standard, discrepancies in data exchange formats exist across different SDMX web services. Secondly, SDMX data requires accompanying metadata for accurate interpretation, yet there has been a lack of robust consideration for a data model capable of concurrently storing both statistical value and its associated metadata. To address these challenges, we propose an scalable data collection framework for SDMX. Structurally, the proposed framework is designed as a flexible and scalable architecture that can be seamlessly extended to target various SDMX open data sources. By deploying dedicated response parsers with standardized in/out interfaces, it can dynamically accommodate a wide range of data sources, providing a scalable solution for diverse statistical data collection. It retrieves data from SDMX open data sources and constructs a integrated SDMX data model within local systems. This model facilitates the retrieval, storage, and management of statistical data while preserving the integrity of the Data Structure Definitions (DSD) as specified by data providers. Additionally, our framework offers advanced data management capabilities, enabling users to efficiently request data CRUD (Collect, Read, Update, and Delete). We validated the functionality and efficacy of the framework by applying it to several prominent SDMX web services.**

Keywords—SDMX, Data management, Data Collection

I. INTRODUCTION

Major national statistical institutes and public authorities, such as the UN (United Nations), the OECD (Organisation for Economic Co-operation and Development), and the IMF (International Monetary Fund), have been publishing high-quality and reliable statistical data that are gathered with significant effort by qualified personnel and are carefully curated, structured, and organized [1, 2]. They have adopted the SDMX (Statistical Data and Metadata Exchange) standard, which describes their statistical data and facilitates publishing them on the Web. SDMX is an international initiative that aims to standardize and modernize the mechanisms and processes for the exchange of statistical data and metadata among organizations [3]. By defining a common language and format for data sharing, SDMX facilitates better interoperability and data integration across various platforms and systems. This standardization is crucial for improving the accessibility and usability of statistical data, thereby enabling more informed decision-making and fostering transparency.

Typically, there are two primary approaches by which end users can collect data from SDMX open data sources. The first approach involves utilizing SDMX queries to retrieve data. However, this approach presents several significant drawbacks. Firstly, there are subtle differences in message exchange formats across various SDMX web services, making it difficult to implement a unified logic for data collection across multiple web services. Secondly, the responses of SDMX queries are normally described with codes to represent attributes of data, aiming to minimize the length of transmission messages. The interpretation of these codes is provided through accompanying metadata, necessitating the inclusion of metadata to accurately interpret the statistical values. Consequently, this method requires an additional process to decode the statistical data, which can introduce complexity and hinder seamless practical use of data. The second approach involves using Python or R-based libraries provided by SDMX web services, which allow users to easily retrieve statistical data into their applications as data objects. However, this approach is not well-suited for handling large datasets, as the object-oriented data retrieval may be constrained by system memory limitations. Also, the use of such libraries often restricts users to the limited information provided by specific SDMX web services, making it difficult to fully leverage the breadth of retained data.

Once end users have successfully retrieved the data, they are confronted with the next obstacle in data storage and management. It is important to recall that, due to the nature of SDMX, data and metadata are managed separately, with metadata being essential for interpreting the statistical data. This separation necessitates a data model capable of storing both the data and the accompanying metadata. However, sufficient consideration has not yet been given to the development of such a data model.

To address these challenges, we propose a scalable SDMX data collection and management framework. A key feature of our framework is its capability to collect statistical data from various SDMX web services using SDMX queries. This functionality is further enhanced by the framework's flexible architecture, which allows for the addition of dedicated response parsers tailored to specific data sources. This structural adaptability enables the seamless integration of multiple data sources, ensuring that the framework can accommodate a wide range of SDMX web services, each with its unique query response format. Once the data is collected, it is converted into an integrated SDMX data model that accommodates metadata, providing semantic context to the data and ensuring comprehensive understanding and utilization. The integrated SDMX data model is critical as it allows for seamless integration and consistent management of disparate data sets, improving data accessibility and usability. Furthermore, the framework supports the efficient management of user-requested statistical data to facilitate updates and maintenance. Ultimately, users benefit from a more efficient and reliable means of accessing, managing, and analyzing statistical data, which improves data accuracy,

979-8-3503-6463-7/24/$31.00 ©2024 IEEE

reduces processing time, and facilitates comprehensive data analysis.

This paper is structured as follows: Section 2 introduces SDMX standard and its supporting IT tools. Section 3 outlines the system architecture and components of our proposed framework, explaining how each component contributes to the overall functionality. Section 4 discusses the implementation of the framework, offering insights into the technical aspects and practical considerations of the system's development. Finally, Section 5 presents the conclusion of the work.

II. Backgrounds

A. *SDMX standard*

The main purpose of the SDMX standard is to address the challenge of creating a common model for the representation of statistical data and metadata [4]. It provides a comprehensive set of guidelines and standards for exchanging statistical information between various entities, such as international organizations and states. SDMX not only standardizes data formats but also introduces metadata as an essential component. This inclusion ensures that the context, quality, and other relevant attributes of the data are adequately represented, making the data more meaningful and usable for various applications. For information exchange, SDMX defines the structure of the content and the message format. SDMX message formats are primarily expressed in two basic forms: SDMX-ML and SDMX-JSON. SDMX-ML uses XML syntax, providing a structured and widely-used format suitable for various data exchange scenarios. On the other hand, SDMX-JSON utilizes JSON syntax, offering a lightweight and easily parseable format that is particularly well-suited for web-based applications and modern data workflows. By using these formats, SDMX facilitates seamless data exchange and integration, promoting interoperability among different systems and organizations. This standardization is essential for enhancing data accessibility, improving data quality, and enabling more efficient and effective use of statistical information.

A key component of the SDMX standard is the Data Structure Definition (DSD). The DSD is a blueprint that defines the structure and organization of dataset. It specifies how data and metadata are to be represented to ensure consistency and interoperability across different systems. The DSD consists of several components:

- *Dimensions* are the fundamental building blocks that define the various aspects of the data. Dimensions are used to categorize data and provide context, such as time periods, geographical areas, or statistical concepts.
- *Attributes* provide additional information about the data, offering further context or qualifying details. They help to enrich the data with more descriptive elements. While the combination of dimensions can identify a single data cell, attributes represent supplementary information about the data.
- *Measures* represent the actual statistical values or metrics being reported. They are the primary data points that users are interested in analyzing.
- *Codelists* are predefined sets of codes used to standardize the representation of dimension values and attributes. They ensure that the data is consistently categorized and easily interpretable.
- *Concepts* define the meaning and semantics of the data elements. They ensure that the data is understood correctly and uniformly across different systems and users.

B. *SDMX IT Tools*

Numerous IT tools [5, 6, 7, 8] have been developed to support the adoption and implementation of the SDMX standard. These tools vary in functionality and cater to different aspects of SDMX data management, ranging from data creation and validation to dissemination and analysis. One of the prominent tools is the SDMX-RI (SDMX Reference Infrastructure), which provides a comprehensive reference implementation of the SDMX standard [9]. It includes components for data and metadata exchange, such as registries, web services, and tools for data validation and transformation. SDMX-RI is widely used by statistical organizations to facilitate the exchange of statistical data and metadata. Another significant tool is the Fusion Metadata Registry (FMR) [10], a robust FMR that enables organizations to manage and disseminate their SDMX data and metadata effectively. It supports the full lifecycle of SDMX data management, including data collection, validation, storage, and dissemination. Fusion Registry offers a user-friendly interface and robust API support for integrating with other systems, making it a versatile tool for statistical data management. Eurostat has developed the Eurostat SDMX Converter [11], which is designed to convert statistical data from various formats into the SDMX standard. This tool supports multiple input formats, including CSV, Excel, and relational databases, facilitating the adoption of the SDMX standard by simplifying the data conversion process.

III. Design of Framework

A. *Overall Architecture*

The proposed framework is designed with three primary objectives, *i*) to facilitate SDMX data collection from various SDMX web services, *ii*) to construct an integrated SDMX data model that combines statistical values and metadata to preserve the semantics, *iii*) to provide comprehensive data management capabilities, including the collection, reading, updating, and deletion of data. To achieve the first objective, the framework's design is built with scalability in mind to allow expansion to accommodate various SDMX web services. Functions that rely on specific SDMX web services, such as parsing query responses, are encapsulated within independent components. These components produce a standardized output, which allows the framework to take on a wide range of SDMX web services. The second objective is to construct an integrated SDMX data model that accurately represents the collected SDMX data. This model combines both the statistical values and the associated metadata, ensuring that the data's context and semantics are preserved. Our system parses the retrieved SDMX query response,

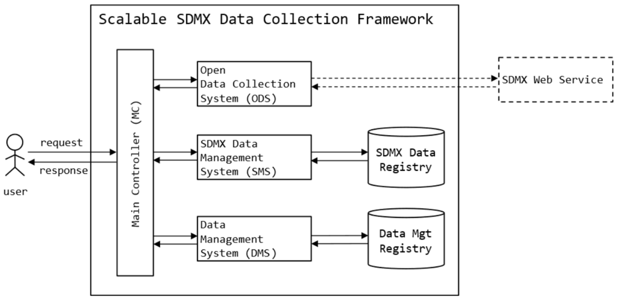
Scalable SDMX Data Collection Framework
Main Controller (MC)
Open Data Collection System (ODS)
SDMX Web Service
SDMX Data Management System (SMS)
SDMX Data Registry
Data Management System (DMS)
Data Mgt Registry
request
response
user

Figure 1. Overall Architecture of the Framework

interprets key elements such as dimensions, attributes, measures, codelists, and concepts, and then organizes these information into an integrated SDMX data model. This model allows users to intuitively understand the data's structure and signification, facilitating better analysis and interpretation. The third objective is to provide comprehensive data management capabilities. This includes supporting read, update, and delete operations on the collected data. The proposed system offers a user-friendly interface and a robust backend system that allows users to manage their datasets effectively. In this framework, statistical data collected through a single SDMX query is managed as a single data unit, while datasets that share the same DSD are organized within a single dataflow. This approach is designed to be efficient from a data management perspective.

The desired functionality is achieved by four subsystems: the Main Controller (MC), the Open Data Collection System (ODS), the SDMX Data Management System (SMS) and the Data Management System (DMS). Each of these subsystems plays a critical role in the overall architecture of the framework, and their integration ensures the system's efficiency and functionality. The following sections provide a detailed description of each subsystem.

B. Main Controller (MC)

The MC manages user interactions and orchestrates the overall control flow within the framework. One key function of the MC is to receive user requests, establish the control flow within the framework based on these requests, and serve as the interface with each component. Users can submit COLLECT, READ, UPDATE, and DELETE requests. The message format of user requests and system responses are described in Table 1. The following segments explain the control flow managed by the MC in response to each type of user request. *i*) COLLECT: the MC first consults the DMS to determine if the SDMX data and structure queries have been executed by a prior user request. If the SDMX data query has already been executed and its data records are available, the MC returns the corresponding data ID to the user. In cases where the metadata is already secured in the framework by the SDMX structure query, the MC forwards only the SDMX data query to the ODS. Conversely, if the structure query has not been executed, the MC transmits both the structure and the data query to the ODS. The ODS subsequently executes the transmitted queries and returns metadata and statistical information to the MC. The MC defines the names for each metadata component and statistical data table, transmitting the metadata along with these table names to the SMS to construct the integrated SDMX data model. After the data model is established, the MC assigns a data ID to the statistical data and forwards it to the SMS to record the data entries. Subsequently, the MC updates the DMS with the new data registration information to ensure proper data management. Once all processes are successfully completed, the MC creates a response message with the relevant data information and returns it to the user. *ii*) READ: after receiving a request containing a data ID, the MC searches the DMS for the relevant information. Based on the retrieved information, the MC then reads the corresponding records from the SMS. Once the records are obtained, the MC generates a response message and delivers it to the user. *iii*) UPDATE: the process is similar but focuses on refreshing existing records with the latest values. When a user submits an UPDATE request with a data ID, the MC retrieves the relevant SDMX data query from the DMS. The ODS processes the query and returns new statistical information to the MC. The subsequent process mirrors the COLLECT procedure. *iv*) DELETE: the MC first verifies the corresponding data information in the DMS. Based on this information, the MC issues a command to the SMS to delete the relevant records. The SMS then proceeds to remove the specified records and returns the outcome of the operation to the MC. Finally, the MC relays the result of the deletion request back to the user. The MC plays a crucial role in handling these user requests. It acts as the primary interface with all system components, ensuring that the appropriate control flow is established based on the nature of the request. Additionally, the MC is responsible for maintaining synchronization between the SMS and the DMS, ensuring that all managed data is consistent and up-to-date.

Table 1. Message format of user request and response

	request	response		
COLLECT	data_name	data_name		
	sdmx_data_query	data_id		
	sdmx_structure_query	dataflow_id		
	update_schedule	update_time		
READ	data_id	data_name		
		data_id		
		update_time		
		obs (list)	meta (list)	concept_name
				code_name
			value	
UPDATE	data_id	data_name		
		data_id		
		dataflow_id		
		dataflow_version		
		update_time		
DELETE	data_id	data_name		
		data_id		
		dataflow_id		
		dataflow_version		
		update_time		

C. Open Data Collection System (ODS)

The primary purpose of the ODS is to execute SDMX queries and parse the responses to extract metadata and statistical information. The system is composed of two components: the SDMX query handler and two types of response parsers, as depicted in Figure 2. The SDMX Query Handler verifies the SDMX data and structure queries and selects the appropriate response parsers to execute them. The response parsers consist of two types: the structure and data query response parsers. The structure response parser executes the structure queries, parses the responses, and extracts the necessary metadata, including code, codelist, concept, concept scheme, dataflow, and DSD. The data query response parser retrieves the statistical values along with the corresponding concept and code combinations and additional data structure information. Each response parser must be specifically tailored for a particular SDMX open data source. While SDMX standards define the components to be included in the

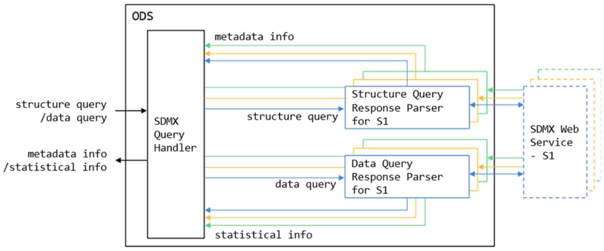

Figure 2. Online SDMX Data Collection System

response, the format and structure of these responses can differ across various SDMX open data sources. Consequently, each response parser requires customized logic to accurately parse the response messages from different data sources. Table 2 describes the interface of the response parsers, which details the components that should be extracted by all response parsers. These components of the metadata and statistical information must be consistently included in the integrated SDMX data model.

Table 2. Input/output Interface specification of the Response Parsers

Input	Output		
SDMX structure query	concept (list)	concept ID	
		concept name	
		concept type	
		concept_scheme ID	
		codelist ID	
	concept_scheme	concept_scheme ID	
		concept_scheme version	
		concept_scheme name	
		concept_scheme agency ID	
		concept_scheme is_final	
	code (list)	code ID	
		code name	
		codelist ID	
	codelist(list)	codelist ID	
		codelist version	
		codelist name	
		codelist agency ID	
	dataflow	dataflow ID	
		dataflow version	
		dataflow name	
	DSD	DSD ID	
		DSD name	
		DSD agency ID	
		DSD is_final	
SDMX data query	observation (list)	meta (list)	concept_ID
			code_ID
		statistical value	
	structure	dataflow ID	
		data source name	

D. SDMX Data Management System(SMS)

The SMS primarily performs CRUD (Create, Read, Update, Delete) operations for statistical data that have been converted into the integrated SDMX data model, as requested by the MC. Before storing statistical data within the SDMX data registry, the SMS is required to construct the integrated SDMX data model based on the provided metadata information. Upon receiving the metadata information and the names of the tables designated for storing each metadata component from the MC, the SMS constructs the integrated SDMX data model as illustrated in Figure 3. Among these tables, the statistics table is created to include columns for the data ID, the statistical values, and all concept IDs defined in the concept table. Additionally, columns corresponding to concept IDs with a "dimension" type are combined with the data ID column to form the primary key (PK) of the table. Once all the metadata tables are constructed and the statistics table is established, the SMS reports back to the MC. Afterward, when the SMS receives the statistical information designated for the statistics table from the MC, it stores the statistical values in the appropriate table. The statistical information corresponds to the 'observation' output interface in Table 2. In the process of recording statistical information, the SMS specifies the code ID as the value for each column that corresponds to a concept ID. It then records the statistical value in the 'value' column that corresponds to the combination of code IDs across all concept ID columns. Read, update, and delete operations in the SMS are all performed based on the data ID. For each operation, the data ID serves as the primary reference, allowing the system to efficiently locate the relevant records within the statistics and metadata tables, ensuring consistent data management across all processes.

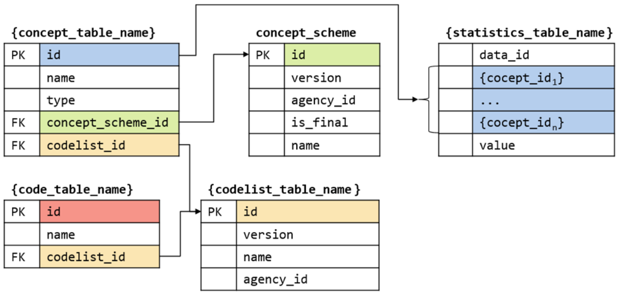

Figure 3. Integrated SDMX Data Model

E. Data Management System (DMS)

The DMS is responsible for ensuring that user requested data are accurately governed and maintained, thereby providing comprehensive and reliable data governance within the framework. Therefore, when user requests that result in actual changes to data records such as collect, update, or delete operations are executed, the registry reflects these effects accordingly. To fulfill this purpose, the DMS organizes its information around three primary entities: Data, Dataflows, and DSD. The relationships between these entities are illustrated in Figure 3, which depicts the relationships essential for maintaining data integrity and consistency.

The Data entity represents a collection of statistical data obtained through a single user request. Consequently, all statistical data within this collection are assigned the same data ID in the integrated SDMX data model. This data ID serves as the operational unit for executing data operations such as read, update, and delete requests in the framework. By organizing data operations around the data ID, the framework enhances both the manageability and usability of the data, ensuring efficient and consistent handling of user-requested datasets. The Dataflow entity represents a collection of data that shares the same context and structure. In this framework, the definition of a dataflow, as provided by the data provider in the SDMX query response, is fully preserved. Consequently, data that belongs to the same dataflow in the SDMX web service will also belong to the same dataflow within this framework. This ensures that all related statistical values are stored in the same data table within the integrated SDMX data model, maintaining consistency and alignment with the original data provider's structure. This allows the framework to provide an accessible pathway for executing read requests on the collected data, ensuring that users can efficiently retrieve the statistical data they have collected. The DSD (Data Structure Definition) entity, as mentioned in the related work, encompasses the structure and semantics of a dataflow.

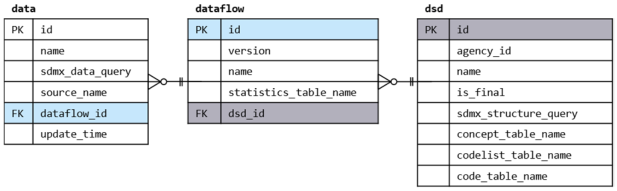

Figure 4. Entity Relation of Data Management Registry

It includes definitions and details regarding the dimensions, attributes, measures, and concepts necessary for interpreting the data. Similar to the approach taken with dataflows, our framework fully incorporates the DSD definitions provided by the data provider. This approach guarantees that the framework maintains a consistent representation of DSDs, even when the same DSD is used across multiple dataflows. Additionally, the DSD entity holds access information about the metadata tables in the integrated SDMX data model. It allows users to access the metadata for data interpretation, so they can retrieve and understand the contextual information required for accurate data analysis. By organizing these three entities, the framework ensures efficient management of the collected data and provides reliable data services to the users.

IV. Implementation and Evaluation

To evaluate the proposed framework, we developed the system using Python and conducted a pilot test targeting two SDMX web services: the OECD Data Explorer and KOSIS (Korean Statistical Information Service). Therefore, we implemented response parsers of the Open Data Collection System for both OECD and KOSIS within the system.

During the testing phase, we performed various data requests to evaluate the functionality of the system, including typical time series data and table-formatted data. Here, we describe the results of COLLECT requests to construct the integrated SDMX data model. We intended to collect the annual real GDP and export data of Korea from the OECD Data Explorer and the annual employment rate from KOSIS. After searching for these data on the respective websites and obtaining the necessary SDMX data queries and SDMX structure queries, we sent COLLECT requests to our system with the following format:

```
COLLECT  request 1:
{
      “data_name” : “annual real gdp of Rep.of Korea”,
      “sdmx_data_query”:
         “https://sdmx.oecd.org/public/rest/data/OECD.SDD.NAD,DSD_NAMAIN
         10@DF_TABLE1_EXPENDITURE,1.0/A.KOR...B1GQ....XDC.L..?dime
         nsionAtObservation=AllDimensions”,
      “sdmx_structure_query”:
         “https://sdmx.oecd.org/public/rest/dataflow/OECD.SDD.NAD/DSD_NAM
         AIN10@DF_TABLE1_EXPENDITURE/1.0?references=all” ,
      “update_schedule”: “”
}

COLLECT request 2:
{
      “data_name” : “annual export of Rep.of Korea”,
      “sdmx_data_query”:
         “https://sdmx.oecd.org/public/rest/data/OECD.SDD.NAD,DSD_NAMAIN
         10@DF_TABLE1_EXPENDITURE,1.0/A.KOR.S1..P6....XDC.V..?dimen
         sionAtObservation=AllDimensions”,
      “sdmx_structure_query”:
         “https://sdmx.oecd.org/public/rest/dataflow/OECD.SDD.NAD/DSD_NAM
         AIN10@DF_TABLE1_EXPENDITURE/1.0?references=all” ,
      “update_schedule”: “”
}

COLLECT request 3:
{
      “data_name” : “annual employment rate of Rep.of Korea”,
      “sdmx_data_query”:
         “https://kosis.kr/openapi/Param/statisticsParameterData.do?method=getList
         &apiKey=?&itmId=T2+&objL1=00+&objL2=0+&objL3=0+&objL4=&ob
         jL5=&objL6=&objL7=&objL8=&format=sdmx&jsonVD=Y&type=Struct
         ureSpecific&prdSe=Y&newEstPrdCnt=100&orgId=101&tblId=DT_1K52
         C01&version=v2_1”,
      “sdmx_structure_query”:
         “https://kosis.kr/openapi/Param/statisticsParameterData.do?method=getList
         &apiKey=?&itmId=T1+T2+&objL1=00+&objL2=0+&objL3=0+&objL4=
         &objL5=&objL6=&objL7=&objL8=&format=sdmx&jsonVD=Y&type=D
         SD&orgId=101&tblId=DT_1K52C01&version=v2_1”
      “update_schedule”: “”
}
```

Our system sequentially generated responses, indicating that the requested "annual real GDP of Korea" was assigned data id as DATA0001 and settled in the DF_TABLE1_EXPENDITURE dataflow. Likewise, the "annual export of Korea" which share a DSD and dataflow with "annual real GDP of Korea" stored in the identical dataflow table. The "annual employment rate of Korea" received data_id DATA0003 and was belong to the DF_101_DT_1K52C01 dataflow.

```
Response 1:
{
      “data_name” : “annual real gdp of Rep.of Korea”,
      “data_id” : “DATA0001”,
      “dataflow_id” : “DSD_NAMAIN10@DF_TABLE1_EXPENDITURE”,
      “update_time” : omitted
}

Response 2:
{
      “data_name” : “annual export of Rep.of Korea”,
      “data_id” : “DATA0002”,
      “dataflow_id” : “DSD_NAMAIN10@DF_TABLE1_EXPENDITURE”,
      “update_time” : omitted
}

Response 3:
{
      “data_name” : “annual employment rate of Rep.of Korea”,
      “data_id” : “DATA0003”,
      “dataflow_id” : “ DF_101_DT_1K52C01”,
      “update_time” : omitted
}
```

Table 3 presents the tables of the integrated SDMX data model constructed from these COLLECT requests. The names of the metadata tables were designated based on the DSD ID, and each metadata table was populated with the corresponding metadata values as defined by the data providers. The statistics tables were denominated by dataflow ID. Additionally, the system updates the data management registry with relevant information. These updates are reflected in Tables 4, 5, and 6, which show the records of the data, dataflow, and DSD tables, respectively. All updated contents in these tables are generated based on the key information parsed from the SDMX query responses, except for the user-defined data names and SDMX query information. Consequently, with a single request containing the SDMX query, users are able to construct a well-structured data model that incorporates both the data structure and the statistical values as defined by the data provider.

Table 3. Constructed components tables of the integrated SDMX data model by the user requests

Table name	Table type	No. records
CONCEPT_SCHEME	concept_scheme	2
CL_DSD_NAMAIN10	codelist	31
CD_DSD_NAMAIN10	code	5153
CT_DSD_NAMAIN10	concept	31
DF_TABLE1_EXPENDITURE	statistics	141
CL_DSD_101_DT_1K52C01	codelist	9
CD_DSD_101_DT_1K52C01	code	13
CT_DSD_101_DT_1K52C01	concept	9
DF_DSD_101_DT_1K52C01	statistics	15

Table 4. Records in Data entity of Data Management Registry by the user requests

id	DATA0001	DATA0002	DATA0003
name	annual real gdp of Rep.of Korea	annual export of Rep.of Korea	annual employment rate of Rep.of Korea
sdmx_data_query	*omitted*	*omitted*	*omitted*
source_name	OECD Data Explorer	OECD Data Explorer	KOSIS
dataflow_id	DSD_NAMAIN10@DF_TABLE1_EXPENDITURE	DSD_NAMAIN10@DF_TABLE1_EXPENDITURE	101_DT_1K52C01
update_time	*omitted*	*omitted*	*omitted*

Table 5. Records in Dataflow entity of Data Management Registry by the user requests

id	DF_TABLE1_EXPENDITURE	101_DT_1K52C01
version	1.0	v2_1
name	Annual GDP and components - expenditure approach	Number of businesses and employees by province, industry, and business category ('06~) (2006~2020)
statistics_table_name	DF_TABLE1_EXPENDITURE	DF_101_DT_1K52C01
dsd_id	DSD_NAMAIN10	DSD_101_DT_1K52C01

Table 6. Records in DSD entity of Data Management Registry by the user requests

id	DSD_NAMAIN10	DSD_101_DT_1K52C01
agency_id	OECD.SDD.NAD	KOSIS
name	National Accounts Main Aggregates10	DSWS Data Structure Definition
is_final	true	true
sdmx_structure_query	*omitted*	*omitted*
concept_table_name	CT_DSD_NAMAIN10	CT_DSD_101_DT_1K52C01
codelist_table_name	CL_DSD_NAMAIN10	CL_DSD_101_DT_1K52C01
code_table_name	CD_DSD_NAMAIN10	CD_DSD_101_DT_1K52C01

Upon executing a READ request for DATA0001, the system responded with a detailed message, which includes the statistical values with specification, ensuring that the user can immediately comprehend the context and significance of the data. This response illustrates the system's capability to generate user-friendly messages that clearly convey the meaning of the statistical values.

```
READ request:
{
      "data_id" : "DATA0001"
}
```

```
READ response:
{
   "data_name": "annual real gdp of Rep.of Korea",
   "data_id": "DATA0001",
   "update_time": omitted,
   "obs": [
      {
        "meta": [
           {"Time period": "2023"},
           {"Frequency of observation": "Annual"},
           {"Reference area": "Korea"},
           {"Institutional sector": "Total economy"},
           {"Counterpart institutional sector": "Total economy"},
           {"Transaction": "Percentage of GDP"},
           {"Price base": "Chain linked volume"},
           {"Transformation": "Non transformed data"},
           {"Table identifier": "Table 0102 – GDP identify from the expenditure
  side"},
           {"Price reference year": "2015"},
           {"Decimal": "One"},
           {"Unit multiplier": "Millions"},
           {"Currency": "Won"},
           {"Confidentiality status": "Free (free for publication)"},
           {"Observation status": "Estimated value"}
        ],
        "value": 1995551400
      },
      ... ,
      {
        "meta": [
           {"Time period": "1953"},
           {"Frequency of observation": "Annual"},
           {"Reference area": "Korea"},
           {"Institutional sector": "Total economy"},
           {"Counterpart institutional sector": "Total economy"},
           {"Transaction": "Percentage of GDP"},
           {"Price base": "Chain linked volume"},
           {"Transformation": "Non transformed data"},
           {"Table identifier": "Table 0102 – GDP identify from the expenditure
  side"},
           {"Price reference year": "1953"},
           {"Decimal": "One"},
           {"Unit multiplier": "Millions"},
           {"Currency": "Won"},
           {"Confidentiality status": "Free (free for publication)"},
           {"Observation status": "Normal value"}
        ],
        "value": 20188200
      }
   ]
}
```

This approach significantly reduces the burden on users compared to directly retrieving data from SDMX web services, as it eliminates the need for parsing response messages and interpreting statistical values manually. By delivering fully interpreted statistical data, the system streamlines the data acquisition process, enhancing usability and efficiency for end users.

V. Conclusion

In this paper, we proposed a scalable data collection and management framework designed to facilitate the collection, storage, and management of SDMX data. The framework addresses technical challenges such as handling varied data exchange formats across different SDMX sources and integrated management of data and metadata. Our evaluation with the OECD Data Explorer and KOSIS demonstrates its capability to dynamically create and manage data models based on user requests. This comprehensive approach ensures that users can handle data with reduced manual effort, high reliability, and improved usability, making it a valuable tool for managing statistical data from diverse sources.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00857, Development of Financial and Economic Digital Twin Platform based on AI and Data)

References

[1] J.Attard, F.Orlandi, S. Scerri, and S.Auer, "A systematic review of open government data initiatives", Government Inf. Quart., vol.32, no.4, pp. 399–418, Oct. 2015

[2] G. Thiry, I. Manolescu, and L. Liberti, "A question answering system for interacting with SDMX database", In: The 6 Natural Language Interfaces for the Web of Data (NLIWOD) Workshop (in conjunction with ISWC), 2020

[3] R. Stahi and P. staab, "History of SDMX", In Measuring the Data Universe, Springer, Cham, pp. 73–83, 2018, DOI: 10.1007/978-3-319-76989-9_11

[4] SDMX(Statistical Data and Metadata eXchange). Jul. 2024. [Online]. Available: https://sdmx.org/

[5] E. Kalampokis, B. Roberts, A. Karamanou, E. Tambouris, K. Tarabanis, "Challenges on developing tools for exploiting linked open data cubes", In 3rd International Workshop on Semantic Statistics (SemStats2015) co-located with the 14th International Semantic Web Conference (ISWC2015), CEURWS, vol.1551, 2015.

[6] S. Fontenay, "sdmxuse: Command to import data fromstatistical agencies using the SDMX standard", The Stata Journal, vol.18, no.4, pp. 863–870, 2018

[7] A. Zancanaro, L. D. Pizzol, R. M. speroni, J.L. Todesco, and F. O. Gauthier, "Publishing Multidimensional Statistical Linked Data", In Proceedings of the Fifth International Conference on Information, Process, and Knowledge Management, pp. 290–304, 2013

[8] R. Stahl and P. Staab, "Working with SDMX," In Measuring the Data Universe, Springer, Cham, pp. 197-210. 2018, DOI: 10.1007/978-3-319-76989-9_13

[9] Eurostat CROS, "SDMX reference infrastructure (SDMX-RI)", Jul. 2024. Available: https://cros.ec.europa.eu/dashboard/sdmx-ri-web-service

[10] D. Araujo, G. Bruno, J. Marcuscci, R. Schmidt, and B. Tissot, "Data science in central banking: applications and tools", IFC Bulletin, no.59, 2023

[11] Eurostat CROS, "The SDMX converter", Jul. 2024. Available: https://cros.ec.europa.eu/dashboard/sdmx-converter