Statistical data governance based on the SDMX

Haïrou-Dine BIAO K.[1,*,†], Emery ASSOGBA[1,†]

[1]Department of Computer Engineering and Telecommunications, EPAC / University of Abomey-Calavi, Abomey-Calavi, Benin

Abstract

Statistics are essential for development of a nation. With arise of technologies such as AI and big data efficient data governance become more and more important to overcome challenges and opportunities evolved by them. Unfortunately, most of databases in our public and private companies and organizations lack interoperability. This work proposes a statistical data governance mechanism based on the Statistical Data and Metadata eXchange (SDMX) standard, designed specifically for statistical data sharing and exchange between organizations.We designed and implemented a statistical database based on SDMX. This system allows more than 10 benin public organizations to be able to produce, publish and share statistical data from various theme. They can express the indicators and levels of disaggregation of these indicators in a flexible way, without having to create a new database.

keywords

statistical data, database, interoperability, SDMX

1. Introduction

Today, digitization and increasing information exchange, statistics play an essential role in the development of nations [1]. To be able to get insight from data, data need to be collected, validated, published and treated. That is made possible by building database and application over these databases to access these data. Unfortunately, the multiplicity of these databases does not allow for efficient data governance. Because it is more difficult to exchange and maintain data between different systems. This work consists in setting up a statistical data governance framework based on the Statistical Data and Metadata eXchange (SDMX) standard, thus enabling other platforms implementing this standard to easily consume the data produced by this database, guaranteeing a high degree of interoperability and reducing the number of databases needed to collect statistical data.

2. Background and state of the art

The rapid advent of information technology has led to a massive explosion of data, creating unprecedented opportunities, but also posing complex governance challenges.

2.1. Data governance

Data governance is defined as "an overall framework within the company for assigning rights and duties to decisions in order to manage data appropriately by as a corporate asset" [2]. It is therefore a set of principles designed to manage the entire data lifecycle, from acquisition to disposal, including use.

Good data governance facilitates exchange and compatibility between different systems and organizations. It thus promotes greater interoperability.

2.2. Interoperability

IEEE defines interoperability as "the ability of two or more systems or components to exchange information and to use the information that has been exchanged" [3]. A specific challenge to interoperability arises from the fact that there is generally no single way of representing information. Thus, the same information content is often represented in different (usually incompatible) ways across different systems and organizations [4]. Data interoperability therefore requires not only the use of standards and metadata, but also the provision of standardized datasets in formats that can be accessed by both humans and machines.

International standards exist for this purpose :

- `Open data`
- `Statistical Data and Metadata eXchange (SDMX)`

2.2.1. Open data

Open data refers to data that is freely available to everyone to use, modify and share without restrictions. For optimal interoperability, data and metadata files must be published in such a way as to be editable by humans and usable by machines, while remaining independent of language, technology and infrastructure. A first step is to make data available via mass downloads in open data formats. There are various fully documented and widely accepted schemas for constructing digital data files, such as CSV, JSON, XML, and GeoJSON, among others [4]. In the context of open data, several catalogs list portals publishing public data [5]. Initiatives include :

- Transnational initiatives such as :
  - World Bank [6],which is one of the promoters of data sources,
  - the databases of the Food and Agriculture Organization (FAO) [7], which cover a wide range of topics related to food security and agriculture. These include :
    - `FAOSTAT`, which provides free access to statistics on food and agriculture (including crop and livestock sub-sectors, etc.);
    - `AQUASTAT`, which gives users access to the main database of country statistics, focusing on water resources, water use and agricultural water management.

International Conference of Information and Communication Technologies of ANSALB (CITA): Security issues in the age of AI, June 27-28, 2024, Cotonou, BENIN

*Corresponding author.

†These authors contributed equally.

dineb90@gmail.com (H. B. K.); emery.assogba@uac.bj (E. ASSOGBA)

- Continental initiatives such as :
  - openAFRICA, an Africa volunteer-driven open data platform that aims to be the largest independent repository of open data on the African continent [8].
- National initiatives such as :
  - Benin open data portal [9];

2.2.2. Statistical Data and Metadata eXchange (SDMX)

SDMX is an international initiative aimed at standardizing and modernizing statistical data and metadata exchange mechanisms. This standard encompasses a data model (the multidimensional data cube), standard vocabularies (content-oriented guidelines), a formal schema definition, and various data serialization formats for building data files and electronic messages for data exchange. In the SDMX ecosystem, data providers can choose between different data serialization formats for sharing datasets, including XML, CSV, JSON, or even EDIFACT [4].

These standards are implemented through new technologies such as application programming interfaces (APIs). APIs facilitate interaction between two different applications so that they can communicate with each other. They act as intermediaries. APIs use the Hypertext Transfer Protocol (HTTP) for cooperation between different programs and web services (REST or SOAP) [10]. They are reusable pieces of software that enable several applications to interact with an information system. They offer machine-to-machine access to data services and provide a standardized means of managing security and errors.

APIs are therefore catalysts for interoperability. However, as their use increases, data security becomes a major concern.

2.3. Data security

Protecting sensitive data is an important part of data governance. It involves implementing measures and protocols to prevent unauthorized access, leakage or manipulation of confidential information.

However, despite efforts to ensure data security, information leaks are still a reality. Due to the rise in API-related vulnerabilities, the Open Web Application Security Project (OWASP), a foundation dedicated to improving software security, has been issuing its list of the top 10 web security vulnerabilities every 2-3 years since 2003. The OWASP foundation's separate classification of the top 10 vulnerabilities for web applications and APIs highlights the divergence between modern APIs and traditional web applications, requiring a tailored security approach.

The OWASP foundation provides a list of the top 10 OWASP API vulnerabilities in 2023 [11] :

- API-1 : Broken Object Level Authorization (BOLA)
- API-2 : Broken Authentication
- API-3 : Broken Object Property Level Authorization
- API-4 : Unrestricted Resource Consumption
- API-5 : Broken Function Level Authorization (BFLA)
- API-6 : Unrestricted Access to Sensitive Business Flows
- API-7 : Server Side Request Forgery
- API-8 : Security Misconfiguration
- API-9 : Improper Inventory Management
- API-10 : Unsafe Consumption of APIs

To prevent such situations, developers need to focus on writing secure code and ensuring that APIs are configured securely. To guarantee API security, it is essential to consider three fundamental pillars of security : confidentiality, integrity, and availability [12].

3. Material and methodology

The aim of this initiative is to ensure interoperability between databases and facilitate the distribution, availability, use and reuse of information. The same applies to data security.

3.1. Material

A set of technological tools including a development environment, programming languages and software tools was used.

3.2. Methodology

There are several stages to the process.

3.2.1. Description of the design of conventional statistical database systems

Setting up a statistical database system involves a series of steps.

The diagram below summarizes the process :

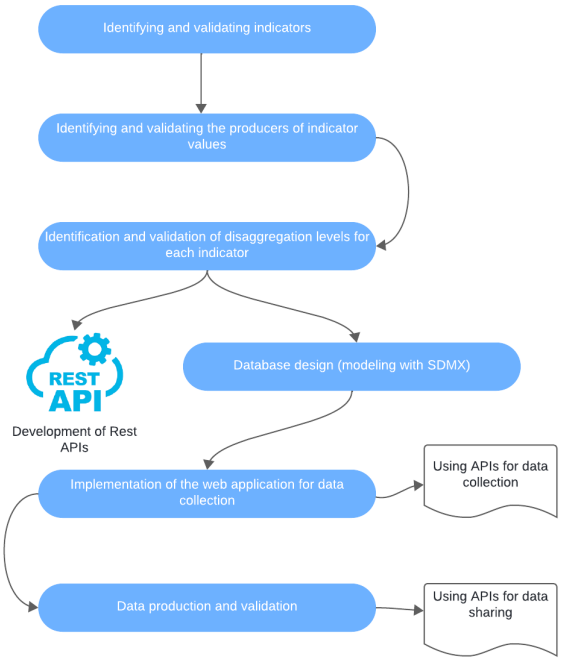

Figure 1: Illustration of the process of setting up a statistical database system.

These are mainly :

- `Identifying and validating indicators` : indicators are quantitative or qualitative measures used to assess the performance or state of a specific domain. They can include statistics such as the unemployment rate, the economic growth rate, the number of new businesses created [13], and so on. This stage involves determining the statistical indicators to be monitored. Indicators are chosen to measure phenomena or evaluate the performance of an action.
- `Identifying and validating the producers of indicator values` : this involves identifying the data sources and the producers responsible for collecting the indicator values. Data producers include the United Nations (UN), the World Bank or the World Health Organization (WHO), etc.
- `Identification and validation of disaggregation levels for each indicator`: this involves determining at what level of detail data will be collected and reported (e.g., by region, by gender, by age group, etc).
- `Database design` : this involves creating a structure for storing indicators and their values in an organized and efficient way.
- `Implementation of the web application for data collection` : this involves developing a web application enabling data producers to submit their data systematically and securely. Tools can be developed to facilitate this stage. The World Bank provides the Survey Solution, a tool for the producing data collection forms. However, Survey Solution requires the installation of a server or the use of the World Bank's demo server and is limited to mobile terminals [14].
- `Data production and validation` : this involves validating and integrating data into the database. To ensure data reliability, several levels of validation are often implemented. In our case, three levels of data validation were necessary before publication.

For example, modeling an observation on the unemployment rate for women in rural areas for the year 2018 in a database can be done through the following parameters :

- `Indicator` : unemployment rate for women in rural areas
- `Disaggregation levels` :
  - Commune : BAN (Banikoara)
  - Department : ALI (Alibori)
- `Observed value` : 2.6% (percentage)
- `Period` : 2018
- `Producer` : the organization or entity responsible for collecting and publishing these statistical data.

However, when it comes to integrating data with any type of indicator and several levels of variable disaggregation, the task quickly becomes arduous because it sometimes requires rebuilding the database. Hence, the need for a standard that takes these complexities into account.

3.2.2. Modeling with SDMX

SDMX is an essential standard for simplifying statistical data modeling and supporting different types of database,

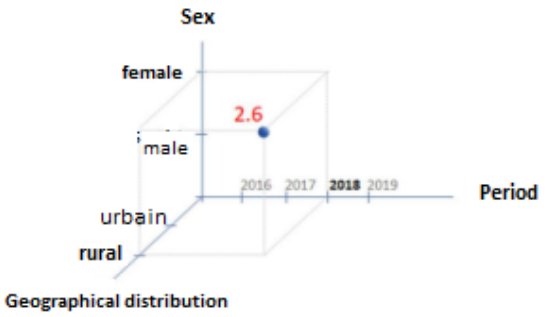

Figure 2: Example of observation representation using the data cube model.

whatever their level of complexity. It uses a multidimensional approach based on the data cube model. The data cube model, also known as the Online Analytical Processing (OLAP) model, is a data modeling method designed to make data analysis and visualization more accessible by presenting it in multidimensional form. Data is organized in cubes, with each dimension representing a distinct aspect of the data.

A multidimensional data cube can be thought of as a model focused on the simultaneous measurement, identification and description of multiple instances of an entity type. A multidimensional dataset consists of several measurement records (observed values) organized along a group of dimensions (e.g., "period," "location," "gender," and "age group") [4]. Using this type of representation, it is possible to identify each individual data point according to its "position" on a coordinate system defined by a common set of dimensions. In addition to measurements and dimensions, the data cube model can also incorporate metadata at the level of the individual data point in the form of attributes. Attributes provide the information needed to correctly interpret individual observations (e.g., an attribute may specify "percentage" as the unit of measurement).

The multidimensional data cube model can support data interoperability across many different systems, independent of their technology platform and internal architecture. What's more, the content of a multidimensional data cube model need not be limited to small datasets. In fact, "with the advent of cloud and big data technologies, data cube infrastructures have become effective instruments for managing earth observation resources and services" [15].

Taking again, the example on the unemployment rate for women in rural areas for the year 2018 its data cube representation can be identified by the following dimensions (figure 2):

- Period = 2018
- Sex = female
- Geographical distribution = rural area
- Observed value = 2.6

In this way, SDMX enables data to be clearly represented by associating measures with dimensions and attributes. SDMX data structures, known as Data Structure Definitions (DSD), describe how data is organized, identifying key dimensions, measures and associated attributes. It also

Table 1
Use of SDMX for improved data representation of an existing platform

Indicator	DEP	COM	SEX	TRANCHE_D_AGE	TYPE_HANDICAP	Period	Value
1	ALI	BAN	TOTAL_F	_T	_T	2024	1

provides standardized terminology for naming commonly used dimensions and attributes, as well as code lists for populating some of these dimensions and attributes. More specifically, a DSD in SDMX describes the structure of a dataset by assigning descriptor concepts to statistical data elements, which include :

- dimensions that form the unique identifier (key) of individual observations ;
- measurement(s) conventionally associated with the concept of "observation value" (OBS_VALUE); and
- attributes that provide more information about a part of the dataset.

In addition, SDMX offers a set of globally agreed DSDs for different application domains, ensuring consistency and interoperability between statistical organizations. [4].

This eliminates the need for a multitude of statistical databases. A single database is enough to federate all an organization's data. The various players can then define the indicators and levels of disaggregation which will be encoded in the database, ready to receive any type of statistical data.

3.2.3. Modeling case with SDMX

Taking the example of the indicator "Total number of support requests for multiple births met" [16], the lack of an effective modeling framework forced the designer to define age ranges as variables, making it difficult to render the data. Using SDMX, the following dimension/attribute levels can be defined :

- SEX with the following code list : F, H, TOTALF, TOTALH
- TRANCHE_D_AGE with the following code list : 0-17-ANS, 18-34-ANS, 35-59-ANS, 60-ANS-PLUS
- TYPE_HANDICAP with the following code list : HMI, HMS, HA, HV, HM, AFH
- DEP with the following code list : ALI, ATA, etc
- COM with the following code list : BAN, NATI, etc

The data representation presented in [16] would amount to this simplified representation (Table 1).

This has the advantage of a more simplified representation and saves storage space eliminating redundancy. The ability to create data structures to define data representation offers great flexibility to data producers, who could define context-dependent data structures for the same indicator.

4. Results

SDMX provides the tools and standards needed to structure open data in a way that maximizes its usefulness and impact.

Its use has enabled us to meet a number of challenges :

- eliminate the multiplicity of databases used to collect and process statistical indicators ;
- put an end to the disparity and scattering of monitoring-evaluation data ;
- an integrated database for storing indicators ;
- effectively operationalize the statistics development strategy ;
- establish a coherent and scalable governance system for statistical data ;
- standardized data ;
- SDMX interoperable APIs, which focus on retrieving metadata and data in XML-JSON-CSV formats. They can used as intermediaries between SDMX-standardized systems or platforms. These platforms include : the .Stat Suite ;
- an environment for producing and disseminating statistical data.

This standardized data is available on a dedicated platform.

5. Conclusion

Data governance, data interoperability and data security are interdependent and essential elements in maximizing the value and minimizing the risks associated with the growing use of data in our society. Implementing the SDMX standard has enabled us to standardize data and obtain interoperable SDMX APIs enabling statistical data to be exchanged between different systems or platforms. The result is an information system based on this standard. There is no longer any need for a multitude of statistical databases ; a single one is sufficient to federate all an organization's data. The various players can then define the indicators and levels of disaggregation which will be encoded in the database, ready to receive any type of statistical data.

References

[1] N. Curien, P.-A. Muet, E. Cohen, M. Didier, G. Bordes, La société de l'information, La Documentation française, 2004.

[2] B. Otto, Organizing data governance: Findings from the telecommunications industry and consequences for large service providers, Communications of the Association for Information Systems 29 (2011) 3.

[3] A. Cooper, Learning analytics interoperability-the big picture in brief, Learning Analytics Community Exchange (2014).

[4] L. G. González Morales, T. Orrell, Data interoperability: A practitioner's guide to joining up data in the development sector. (2018).

[5] M. TRAORE, Les banques de données environnementales (????).

[6] W. Bank, World bank open data, 2024. URL: https://data.worldbank.org/.

[7] Food, A. O. of the United Nations, Statistiques, 2024. URL: https://www.fao.org/statistics/fr.

[8] openAFRICA, Africa's largest volunteer driven open data platform, 2024. URL: https://open.africa/.

[9] Bénin, Un coup d'oeil sur les données du benin, 2024. URL: https://benin.opendataforafrica.org/.
[10] A. Soni, V. Ranga, Api features individualizing of web services: Rest and soap, International Journal of Innovative Technology and Exploring Engineering 8 (2019) 664–671.
[11] D. Timsina, L. Decker, Securing the next generation of digital infrastructure: The importance of protecting modern apis (2023).
[12] H. Asemi, A study on api security pentesting (2023).
[13] G. Zakhidov, Economic indicators: tools for analyzing market trends and predicting future performance, International Multidisciplinary Journal of Universal Scientific Prospectives 2 (2024) 23–29.
[14] L. J. Young, G. Carletto, G. Márquez, D. A. Rozkrut, S. Stefanou, The production of official agricultural statistics in 2040: What does the future hold?, Statistical Journal of the IAOS 40 (2024) 203–210.
[15] S. Nativi, P. Mazzetti, M. Craglia, A view-based model of data-cube to support big earth data systems interoperability, Big Earth Data 1 (2017) 75–99.
[16] SiDoFFe-NG, Statistiques detaillees du domaine protection sociale et solidarite nationale, 2024. URL: https://2019a2024.sidoffe-ng.social.gouv.bj/sidoffepublic/stats/details/pssn.