Modernizing data management at the US Bureau of Labor Statistics

Daniel W. Gillman and Clayton Waring[1]

Abstract

The US Bureau of Labor Statistics (BLS) is undertaking initiatives to improve its data and metadata systems. Planning for the replacement of the public facing LABSTAT data query system and efforts within the Office of Productivity and Technology to combine multiple production systems within a single cross-divisional database platform are examples. BLS views time series data as a combination of three elemental components found in every time series. A measure element; a person, places, and things element; and a time element are the components. The authors turned this basic approach into a formal conceptual model represented in UML (Unified Modeling Language). The UML model describes a flexible multi-dimensional data structure, of which time series are a kind, and supports any kind of query into the data. The Office of Productivity and Technology adopted the model, and it is guiding their approach moving forward. The Financial Industry Business Ontology project under the Object Management Group and the Data Documentation Initiative Cross-Domain Integration (DDI-CDI) development project also adopted the model. In this paper we describe the time series formulation and the UML conceptual model. Then, the design of the OPT system and its features are described. In doing so, we provide a thorough understanding of the structure of time series.

Keywords

Multi-dimensional data, time series, metadata, measures

Introduction

The US Bureau of Labor Statistics (BLS) is undertaking initiatives to improve the way it manages its data and metadata systems. Two examples include planning for the replacement of its public facing LABSTAT data query system and efforts within its Office of Productivity and Technology to combine multiple production systems within a single cross-divisional database platform.

Within these projects, BLS views time series data as a combination of three mutually exclusive elemental components, which are found in every time series. They include a measure element; a person, place, and thing element; and a time element. The authors turned this basic approach into a more formal conceptual model represented in the Unified Modeling Language under the Object Management Group[2] (OMG) [OMG, 2017a]. The UML model describes multi-dimensional data, of which time series are a kind, and is very flexible in that it supports any kind of query into the data.

The Office of Productivity adopted the model, and it is guiding their approach moving forward. The OMG Financial Industry Business Ontology – Indices and Indicators (FIBO) [OMG, 2017b] standard and the Data Documentation Initiative (DDI) Cross-Domain Integration (DDI-CDI) development project under the DDI Alliance[3] [DDI Alliance, 2020b] also adopted the model.

 Gillman, D. W. and Waring, Clayton (2023) Modernizing data management at the US Bureau of Labor Statistics, IASSIST Quarterly 47(1), pp. 1-15. DOI: https://doi.org/10.29173/iq1038

In this paper we describe the elemental components of time series and the UML conceptual model derived from them. Then, the design of the OPT system and its features are described. In doing so, we provide a thorough understanding of the structure of time series.

History

There have been many attempts to describe multi-dimensional data over the years. In all these efforts, the structure of the data is elemental. It defines the semantics (i.e., the meaning) of the data and the end points for beginning a query.

In his PhD dissertation in 1973, Sundgren [Sundgren, 1973] developed a mathematical treatment of multi-dimensional structures called boxes. A box has dimensions and a measure, and the cells defined by the elements of the dimensions store the values. Additivity of cells, used to aggregate data and reduce the number of dimensions, is inherent in the approach. The main contribution is the idea that a table is just a presentation format for boxes.

After the development of the relational model by Codd [Codd, 1970], data warehouses and the online analytical processing (OLAP) model followed, pioneered by E. Codd, S. Codd, and Salley [Codd, et al., 1993]. OLAP evolved as a counterpart to online transaction processing (OLTP), the method for handling business transactions, to aid business analyses. As with Sundgren, the focus of the OLAP approach was query, not semantics.

The STORM (Statistical Object Representation Model) model for describing multi-dimensional data in was introduced in 1990 [Rafanelli and Shoshani, 1990]. This approach recognized that a table presented an arbitrary ordering of the dimensions. The STORM system provided the ability to manipulate the dimensions so a new table with a new ordering of the dimensions could be produced. It relied on additivity to facilitate the necessary transformations.

In 2000, van Bracht [Van Bracht, et al., 2000, and Van Bracht, 2004] resurrected Sundgren's box model and referred to the structure as an n-cube. These n-cubes include a measure associated with each of the cells. This paper cemented the distinction between n-cubes and tables in the statistical community. Additivity of data and projections (i.e., aggregations and eliminating dimensions) were an important focus.

The first, simple, standard in the DDI suite is DDI-Codebook, first released in 2000. As of this writing, the current version is 2.5. [DDI Alliance, 2012]. This standard contains a mechanical format description of a table. It includes the dimensions and measures in the description of each table.

The Statistical Data and Metadata eXchange (SDMX) specification was released [SDMX, 2004] in 2004, and version 3.0 of this standard was released in 2021 [SDMX, 2021]. SDMX explicitly implements the idea of an n-cube. It is possible to treat time itself as a dimension, and SDMX incorporates the ability to include values from more than one measure in each cell. The focus of SDMX is the exchange of multi-dimensional data with metadata attached, mostly the dimensions. The World Wide Web Consortium (W3C)[4] issued their Data Cube Vocabulary [W3C, 2010], a version of the SDMX data model rendered in RDF.

The standard in the DDI suite based on the statistical lifecycle is DDI-Lifecycle (DDI-L), first released in 2008. The latest is version 3.3, released in 2020 [DDI Alliance, 2020a]. This includes the n-cube idea

and has similar limitations and restrictions as in SDMX. The focus on the statistical lifecycle distinguishes DDI-L from SDMX. DDI-L also supports metadata reuse.

In 2012 the UNECE released the Generic Statistical Information Model (GSIM). The latest version, released in 2019, is version 1.2 [UNECE, 2019]. GSIM is a conceptual model of the information needed to describe statistical data production, and it includes multi-dimensional data as n-cubes. It does not explicitly include time series. The treatment of n-cubes is like that found in SDMX and DDI-L.

The new DDI Cross-Domain Integration (DDI-CDI) standard, for release in early 2023, includes a full description of n-cubes [DDI Alliance, 2020b]. Multi-dimensional data, of which n-cubes and time series are kinds, are among the four major data structure types that DDI-CDI describes. The model presented in this paper is the basis for that descriptive capability. The OMG FIBO standard included the model presented here as well.

Background

Our problem is to find a simple and intuitive way to store and organize statistical data with the goal of making it easy to find and use the data. The challenge is to reduce the overall complexity of the problem by uncovering underlying principles from which to build. The semantic approach we propose contains these principles and supports storing, organizing, and querying data based on their meaning, not their structure. These principles are simple and lead to a natural design. We avoid complexity and rigidity by building from the principles.

Einstein is famously attributed as to having said, “Everything should be made as simple as possible, but no simpler.” At BLS, the LABSTAT system is the main dissemination tool for BLS time series, and, paradoxically, it is both too simple and complex. It is too simple because it does not support a wide variety of queries (functionality). For example, it is not possible to find all data BLS has about nurses, hospitals, or North Carolina through a simple query. But LABSTAT is also too complex in that its underlying specialized design is based on diverse, siloed principles that do not easily scale or support flexibility. This complexity results in a rigid design that is difficult to improve.

To advance simplicity and increase functionality, we begin with the notion that storing and organizing things (e.g., data) is a simple and intuitive concept. People in their routine lives store and organize dishes in kitchen cabinets, clothes in bedroom dressers, books on shelves, and cars in garages. Fundamentally, storing and organizing statistical data should not be any different than these examples. One thing cabinets, dressers, bookshelves, and garages all have in common is that they are designed specifically to accommodate the common definitional attributes and characteristics that define dishes, clothes, books, and cars. In other words, the method of storage and organization depends on the items being stored and organized. So, a first step in storing and organizing data is to understand the common definitional attributes and characteristics of data.

A time series is a set of numeric values that represents the change over time of measurable aspects of people, places, and things. Under the approach presented here, statistical data is stored and organized around the elemental components: measures; people, places, and things; and time.

The first elemental component is people, places, and things. This refers to groups that share a set of commonalities, such as nurses in North Carolina, shirts sold in Pittsburgh, U.S. manufacturing capital equipment, men in New England, heads of household over the age of 40, etc. The next component is

measures, which generally refer to quantitative, or numeric, variables. These include things like consumer price index, unemployment rate, value of GDP, crop yields, etc. The time component is simply the time stamp that identifies the applicable time reference for the observation.

These three common elemental components connect all statistical data through a shared understanding of the meaning of data. For example, the price of shirts in Pittsburgh in the first quarter of 2016 and the unemployment rate of men in New England in 2020 both include measure; people, places, and things; and time components. So, one advantage of this approach is it allows seemingly diverse statistical data to be stored, organized, and queried within a single, unified framework.

Semantic Approach

Rather than taking the usual structural approach to describing multi-dimensional cubes, time series, and their associated measures, we adopted a semantic one based upon the "Measures / People-Places-Things / Time" model. By this we mean the focus is on the meaning of the data rather than on some predefined structure. The structure follows naturally from the semantics.

People-Places-Things

Consider the description "Nurses working in Hospitals in North Carolina". This is an example of a PPT (People-Places-Things) description. Technically, we call this a universe. It can be used to describe the interpretation of a cell in an n-cube or the focus of a time series without knowing anything about a particular measure.

All statistical data have a geographic location to which data apply. In our example it is the state of North Carolina. The physical location of hospitals as a place of employment necessitates the geographic component. This is a consideration in all cases. Sometimes the area is just implied.

The primary component of a universe is called a unit type. This addresses the basic units of analysis in the data. In other words, the unit type comprises the things in the world the data describe. Unit types may be persons, households, business establishments, animals, commodities, economic constructs (output, labor, capital, and materials), or events (e.g., marriage, education, or hospitalization), among others.

Some unit types can be usefully partitioned. A business establishment can be partitioned into output, labor, capital, and materials. A household can be partitioned into liabilities and assets, and assets can be further partitioned into financial and material. These act as unit types themselves.

Unit types are specialized into universes by applying the other categories in a PPT description. However, starting with the description of a universe, the unit type is not necessarily implied. Our example, "Nurses working in hospitals in North Carolina" is a case in point. There are at least two ways to unpack and understand this description, because different unit types might apply.

Starting with "nurses", nursing by itself is an occupation, so "nurses" is a combination of people (a unit type) and an occupation. This is one way to interpret the example phrase. Further, in this case, we only care about those nurses working in hospitals from the nurses' perspective, as nurses can work in many other settings.

On the other hand, we could start with hospitals, an industry, and recognize that hospitals are a kind of business establishment (another unit type). In this interpretation, we care about the nurses working

in hospitals from the hospitals' perspective. So, in this case, "nurses" refers to a category of employee in hospitals (along with others), and hospitals are business establishments specialized to an identified industry.

In both perspectives we have a different unit type – people for the nurse perspective, and business establishments for the hospital perspective. Therefore, we have 2 potential interpretations of the universe:

1. People working as nurses in hospitals in North Carolina.
2. Hospitals (which are business establishments) employing nurses in North Carolina.

The choice comes down to what the basic unit of analysis is designed to be, and this follows from the design and meaning of the underlying data.

In each case, we have unpacked the description to identify the unit type and the categories used to specialize that unit type into a universe. Even though the words used to name the universe in the two interpretations are similar, the underlying meaning is different, because the starting point, the unit type, is different.

Measures

Measures represent the basic notions of size and scale. Therefore, measures can be thought of as pairing concepts of size or scale to numeric values. For example, the size of a house may be equal to 2000 ft^2 in January 2022. This is subject to direct comparisons. For example, the number of square feet can used to compare with the size of a different house. Alternatively, the square footage of a house can be compared over a time interval to see if the size is increasing or decreasing.

The numeric values assigned to a measure are called measurements. So, the idea of the size of a house is a measure while the number of square feet of a particular house is a measurement. And, within this context, certain rules must be adhered to. For example, the size of a house cannot be quantified using measurements such as the miles per hour or degrees of temperature. Size and scale may also be associated with economic concepts such as prices, employment, sales, wages, and others. For example, the price of a gallon of milk may be equal to $4.50 USD.

Separating the measure from the numeric measurement is useful for several reasons. First, it permits the description of the measure to be independent of the description of the measurement. For example, any number of descriptors can be added to a measure, such as price. The price can be described as a consumer price excluding sales promotions and prior to the addition of applicable sales taxes. Similarly, other descriptors can be added to the numeric measurement. As above, $4.50 can be described in units of current U.S. dollars.

Another reason for separating the measure from the measurement is their relationship to time. The concept of a measure tends to be reliably steady; however, the numeric measurement may change over time. For example, the concept of a price remains the same, however, the actual price of an item, say a gallon of milk, changes quite frequently. Consequently, a measurement is associated with a particular observation at a point or interval in time. In other words, the measure refers to a definition associated with something like price, whereas the measurement refers to a specific price observed at a known time.

The last reason for separating the measure from the measurement is it reveals how multiple measurements can be assigned to the same measure for one reference time. For example, the price of an item on some date can be given in USD, Pesos, or Euros.

Within the model items are grouped by their similarities and separated by their differences. So, two measures of sales have similarities, but there may be differences that lead to different values for the observations. A measure of sales could be gross sales or value-added sales. The measures might also differ by the survey sample, the survey frequency, calculation methodology, or any adjustments that are made to the results. In these cases, the description of the measure is used to differentiate the understanding of outcomes for the purpose of comparisons. The proper description of a measure ensures the correct understanding of the measurements, reducing the risk of misinterpretation.

Time

Time, as might be expected, refers to our innate understanding of the sequence of events we observe in our daily lives. It is referred to in standardized units such as year, months, and days to enhance communication. Further, time is referenced as a point in time or as an interval of time (with begin and end points). Measures can refer to either, whereas time series refers to intervals.

Within our model time plays several important roles. First, time is used to compare measurements. Measurements of wages, prices, and employment may fluctuate over time, and this notion of change over time is what characterizes measures. Changes of values of numeric measurements over time are what economic analyses investigate. Measures may have attribute qualifiers assigned to them, such as seasonally adjusted, but these attributes do not change over time. In other words, when seasonally adjusted employment increases, it is understood that employment is the concept that is increasing. In this way, 'seasonally adjusted' is the type or kind of employment that is increasing. As such, seasonal adjustment is not considered to be a measure in and by itself. So, one of the things that differentiates measures and the attributes assigned to them is their relationship to time.

Second, time is used to classify measures based how they relate to the passing of time. This is perhaps best illustrated by considering what would happen if time were to be held still as in a science fiction movie. With time held still certain measures would continue to exist, such as the number of employees at a firm or the value of the firm inventories. However, other measures such as the number of average weekly hours worked at the firm would cease to exist. It is impossible to work for an hour while time is held still. Measures that exist at an instance of time are broadly referred to as stock measures while measures that require the passing of time are referred to as flow measures. However, measures of central tendency and dispersion, such as averages and standard deviation, apply to both stock and flow measures. These represent yet another class or type of measure.

The third role time plays in the model is identifying the time stamp for the measure observations. The time stamp identifies the sequence of the observations. Furthermore, the time stamps determine if the observations have regular intervals, with equal time intervals between the observations, or not. However, the meaning or interpretation of the timestamp is slightly different for stock and flow measures. A timestamp of 2021 represents an interval, typically from January 1st to December 31st. Interval timestamps naturally fit with flow measure such as work hours and value of shipments. The measurement of value of shipments for 2021 would presumably be the sum of all shipments from the beginning to the end 2021. However, for stock measures, interval timestamps can be problematic. For example, consider the measurement of employment at a firm for 2021. Employment cannot be

measured as the sum of all employment from the beginning to the end of the year. More likely, the employment measurement may be some type of average over the year or the employment level on a specific day of the year. In each case additional information beyond a simple timestamp may be needed to provide clarity to the meaning of the measure and its relationship to time.

UML Model

UML is a widely used modeling language. It is built for object-oriented analysis and design (OOAD), a systems design paradigm. Within that framework, several modeling techniques are defined and used, including class diagram, which show how entities (called classes in OOAD) are described and related. Class diagrams depict the entities, attributes, and relationships in a visual framework. For the uninitiated, there exist many online tutorials on how to construct and read UML class diagrams[5].

UML class diagrams are useful outside the specifics of OOAD, and that is our approach in this paper. Based on specificity, there are three levels of class diagram models for systems development:

- Conceptual
- Logical
- Physical

These levels are explained in simple language in tutorials on the Web[6].

In this paper, we are interested in the least specific of these, the conceptual model. A conceptual model is designed for human communication. It is used to communicate the essential requirements for some system, and these correspond to the entities, attributes, and relationships in the model. A system built to satisfy those essential requirements is said to conform to the conceptual model.

Logical models are used to establish and further refine the requirements for a system. The physical model contains all the requirements in a logical model and corresponds to the schema for an existing system.

In this section, we describe the conceptual UML model that follows from the PPT, measures, and time components in our semantic approach. The model is presented in four class diagrams, Figures 1-4. The colors in the boxes depicting classes indicate which of the components are addressed with each class: dark blue for PPT, light blue for Measures, and brown for Time and Multi-dimensional Structures.

PPT Model for People, Places and Things

As described above, there are three main concerns in the construction of our conceptual model (for simplicity, “model”): PPT, measures, and time. In this section, we describe how PPTs are modeled and why the model is structured as it is. The PPT construct translates into a Universe in the BLS model. A universe is all the specialized units that apply to some data. Our previous example of nurses working in hospitals in North Carolina is typical.

The first consideration is the fundamental units we are measuring, e.g., people, households, establishments. For nurses working in hospitals in North Carolina, placing nurses first in this formulation makes us think of people. If we had said hospitals employing nurses in North Carolina, our

focus shifts to hospitals. The main point is that word order really does not matter. Let us assume we mean people as the fundamental units, which we refer to as *UnitType*.

Nurse, hospital, and North Carolina are each a *Category* in our model. The set of *Categories* from which nurse, hospital, or North Carolina is drawn is a *Dimension*. For example, the North American Industry Classification System[7] (NAICS) is divided into levels, each expressing different levels of detail. Hospital is the category labelled 622 in the subsector level of NAICS. So, the *Dimension* in this case is a list of subsectors, with hospital listed as a *Category* within. There are similar considerations for nurse in the Standard Occupational Classification[8] (SOC) and North Carolina in a list of US states such as the US Postal State Codes.

The SOC is a hierarchical listing of occupations, NAICS is a hierarchical listing of industries, and Postal codes enumerate the US states. Each of these general considerations – occupation, industry, and states of the US – is a *Characteristic*.

Each *Category* specializes the *UnitType* people, for instance people residing in North Carolina, people employed by hospitals, or people working as nurses. The combination of the three *Categories* further specializes people in this case. The result is a *Universe*. It is a specialization of a *UnitType*.

Some unit types can be meaningfully broken into parts, and this is how the *UnitTypePartition* is used. Each element of a partition can be a *UnitType* itself, a *FacetedUnitType*. For example, establishments can be subdivided into output, labor, capital, and material. Households can be subdivided into liabilities and assets. Each could be used as a unit type.

A combination of *Categories*, each from one of the *Dimensions* corresponding to each one of the *Characteristics*, all applied to a *UnitType*, forms a *Universe*, which specifies a subset of the *UnitType*. A *ScopedMeasure* is the result of the association of a *Universe* with a *QualifiedMeasure*, which are discussed in the Measures section below. A *Universe* describes the units (or objects) to which a *ScopedMeasure* applies. All these ideas just discussed are illustrated in Figure 1 below.

Measures Model

Previously, we said measures and measurements are different, and among measures there are similarities and differences. So, we treat measures and measurements separately, and we specify several layers for measures. See Figure 2 below to view the BLS model for measures, which we describe in the following.

In our formulation, a *Measure* is a quantitative variable. It has a topic, or concept, associated with it, that groups broadly similar measures, but this is outside the scope of our model. One such concept might be wages, which can be specialized into various measures, such as weekly wages, hourly wages, average weekly wages, median weekly wages, etc.

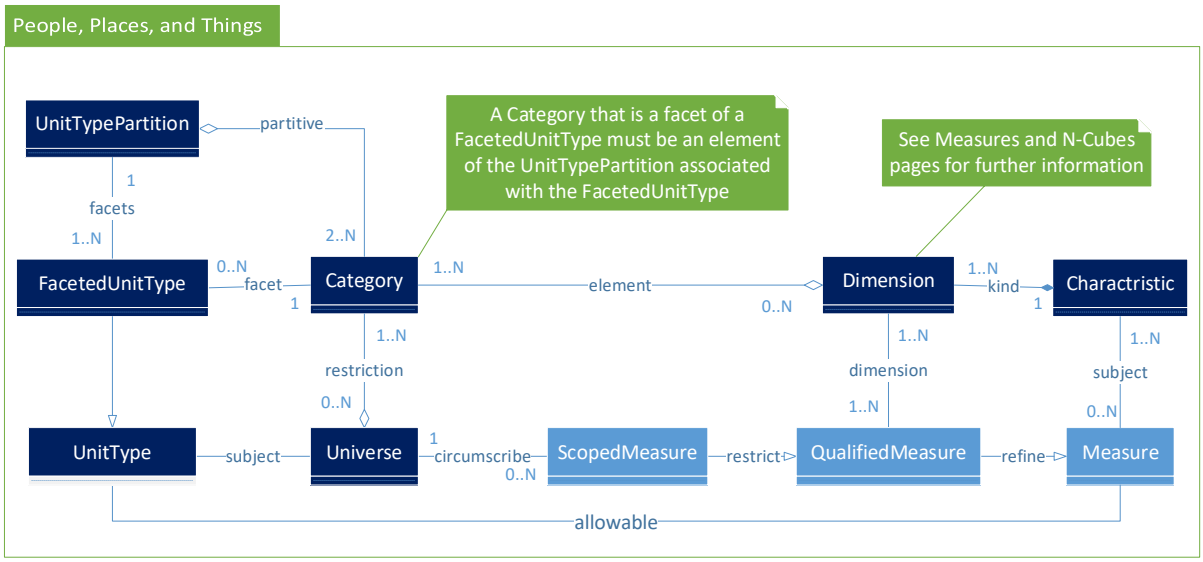

Figure 1: PPT Model

A *Measure* has a *DataTypeFamily* associated with it. This is a broad characterization of what one can do with the associated data. *Measures* are quantitative variables, and the main quantitative types are interval and ratio. Rates, percentages, and totals are typically ratio data. Indexes are interval. The difference is whether a value of 0 means absence of the quantity: ratio is yes; interval is no. The *UnitOfMeasure* is the specific way quantities are named, such as dollars for expenditures or wages.

There are many ways a *Measure* can be made more specific. For instance, "average weekly wages" is produced in at least 2 ways at the Bureau of Labor Statistics: 1) from the Current Employment Statistics[9] (CES) survey estimates of average hourly wages and weekly hours worked; and 2) from the Quarterly Census of Employment and Wages[10] (QCEW) estimate of average quarterly wages divided by 13. Both measures fall under the same grouping: average weekly wages. We account for these differences by *ProductionMethod*. *DisclosureAvoidance* or *Adjustment*, such as seasonal adjustment, are additional ways in which a measure can be differentiated.

The resulting specialization of a *Measure* is called a *QualifiedMeasure*. Here, the *Dimensions* associated with the generic *Measure* are linked. For example, the set of sectors in NAICS is a specific instance of the *Characteristic* industry.

The combination of a *Universe* and *QualifiedMeasure* is a *ScopedMeasure*. This is the specialization of *Measure* that we associate data with, and this done through *Measurement*. Each *ScopedMeasure* may have many *Measurements*. Here, the Universe tells us the relevant set of things (units or objects) in the world to which the data apply. A *Measurement* has a *Datum* and a *TimeStamp* (usually a reference date and a release date). Sometimes, the data associated with a *ScopedMeasure* for some reference date is revised (e.g., the total jobs added to the US economy in February 2021 was revised upward from 468,000 to 536,000 in June). For this reason, more than one *Datum* may be associated with a single reference date. *Revision* provides the reasons for why a number might be changed. The attribute vintage in the class *Datum* is incremented by one for each new revision. This number allows researchers to put together series of data based on which revision is relevant for investigation.

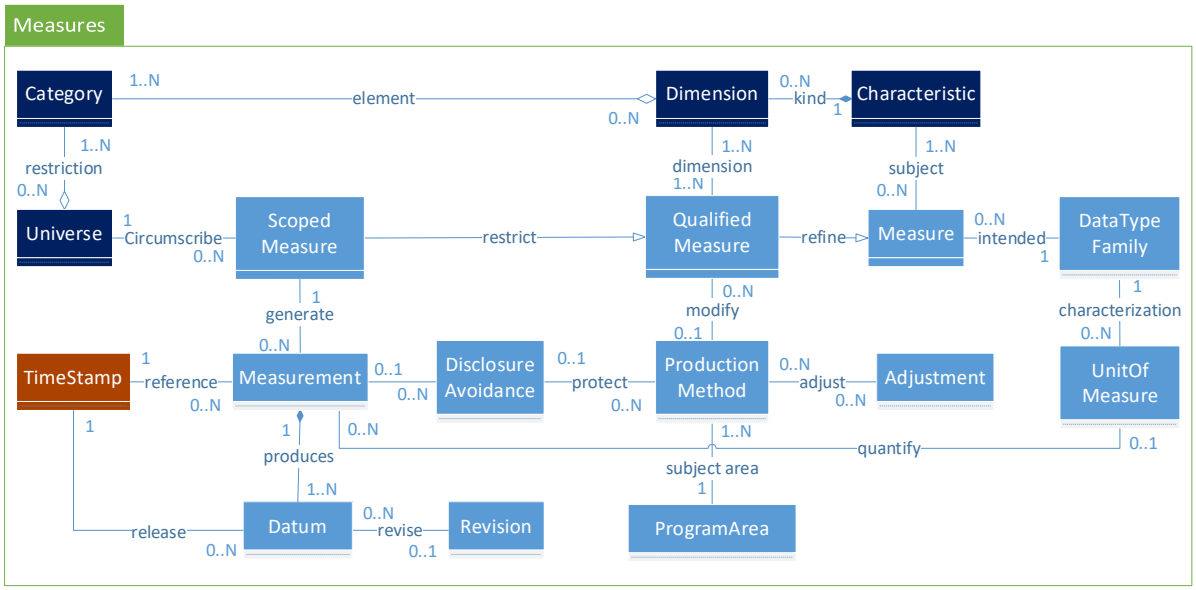

Figure 2: Measures Model

Multi-dimensional Structures

There are 2 ways to use a *ScopedMeasure*: to express how *Measurements* change over time in a *TimeSeries*; or as an expression of related *Measurements* (with other *ScopedMeasures* under a single *QualifiedMeasure*) in an *N-Cube* for a single reference period. The models for each follow below, and entities with the same name in these models and the ones above are the same entity. Figure 3 depicts the structure of time series, and Figure 4 depicts the structure for an n-cube.

These models describe the full range of multi-dimensional data produced by BLS, with one caveat. Some data produced by BLS are in the form of time series, such as those from the Quarterly Census of Employment and Wages (QCEW), but some of them are not strictly time series. QCEW does not redefine historical data to align with changes to NAICS. So QCEW aggregates (e.g., county employment) are time series, but the detailed industry data are only time series in a short-term sense.

In Figure 4, we distinguish between the structure of an n-cube (a *StructuralN-Cube*) from an n-cube with data applied (an *N-Cube*). The structure is just all the *Dimensions* and their underlying *Categories*.

Figure 3 includes an entity called *MeasurementGroup*. It allows us to lump more than one *ScopedMeasure* into a *TimeSeries*. This violates the definition of a time series, but BLS uses it to describe inter-related series more easily. An example is the combination of the US Unemployment Rate and the percentage point change from the previous reference date. SDMX contains this feature as well.

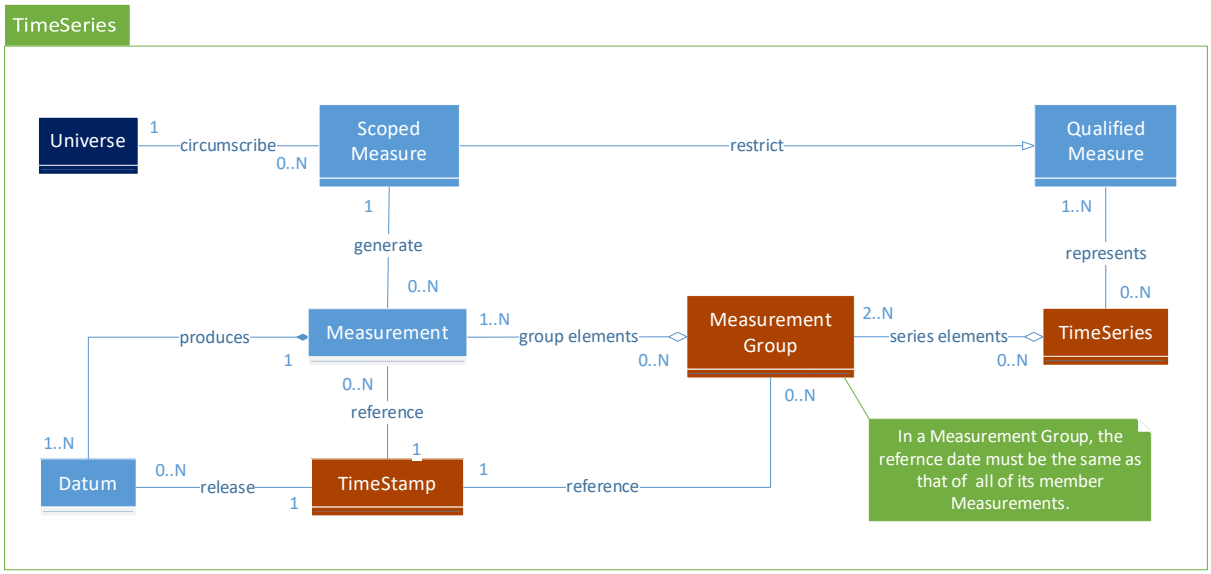

Figure 3: Time Series Model

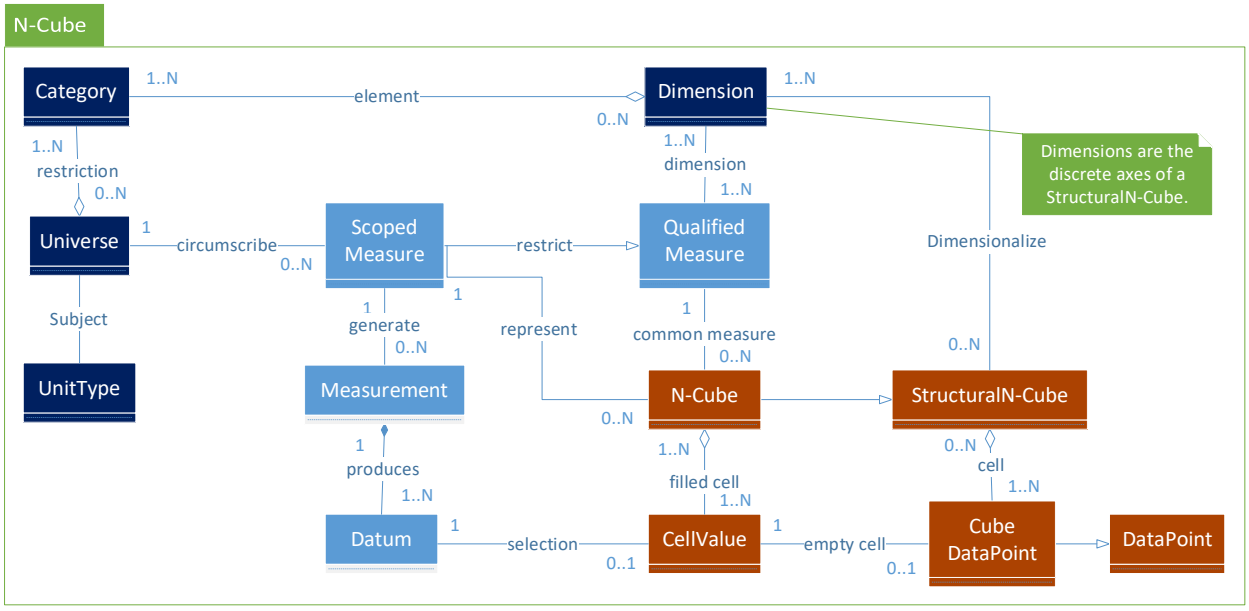

Figure 4: N-Cube Model

Applications

The BLS Office of Productivity and Technology (OPT), unlike other program offices in BLS, does not process survey data directly but rather collects data from several sources. These data sources include other BLS programs and outside sources such the Bureau of Economic Analysis (BEA), the Census Bureau, and others. These put OPT into a situation where it is both a data user acquiring processed data from other sources and a data disseminator producing statistical data for public consumption. The problem facing OPT was a common one: most data sources use their own idiosyncratic method to organize data. This leads to a siloed approach that inhibits interoperability within a statistical production system. The Measures / People-Places-Things / Time approach alleviated much of the

siloed data issue and allowed OPT to begin combining previously incongruent datasets into a uniform structure for the purpose of storing, managing, and querying data. All the statistical data in the system share a common definition of data that is expressed by the unifying semantic model presented here. Currently, OPT is combining two cross-division productivity production systems into a single system.

In the example in Table 1 below, the BLS 2020 index value of labor productivity for workers in North Carolina demonstrates how the model is used to organize and structure metadata elements within a statistical production system.

Each time series is identified by the unique combination of the field values associated with the people, places and things element and the measure element. The data value within the time series is then uniquely identified by the addition of the time stamp and vintage fields. The system also allows for additional metadata elements. The choice of the fields and their associated field values will differ from one dataset to another.

Model		Field	Field Value	Description
Scoped Measure (Time Series)	People, Places, and Things (Universe)	UnitType	Establishment	Sets the UnitType as establishment with added specificity of private nonfarm business establishment. UnitTypePartion set as labor.
		EstablishmentType	Private Nonfarm Business	
		UnitTypePartition	Labor	
		Industry	113_81	Sets the category of industry to the characteristics of NAICS code 113_81. Sets the category of geography to the characteristic North Carolina
		Geography	North Carolina	
	Measure	MeasureLabel	Labor Productivity	Sets the principal measure to labor productivity. Labor productivity is output per unit where unit is set to the UnitTypePartition of labor.
		Measure	Output Per Unit	
		MeasureProgramArea	BLS Office of Productivity	Sets various measure attributes which qualifies the measure.
		MeasureAdjustment	Not Seasonally Adjusted	
		MeasureMethodology	Value Added	

Model		Field	Field Value	Description
		MeasureType	Index	
		MeasureUnit	2007=100	
		MeasureDisclosure	Public Access	
		MeasurementDatum	110.823	The assigned value of the labor productivity index.
Time	Time	TimeStamp	2020	Sets the TimeStamp to 2020 and sets the vintage to the publication date of May 27, 2021, for the purpose of tracking revisions over time.
		Vintage	5/27/2021	

Table 1: OPT Example

As mentioned in the Introduction, the model described herein was adopted by two standardization efforts, FIBO and DDI-CDI. As of this writing, each of these efforts is near to publishing the first version of its standard. The DDI-CDI effort is expected to produce its first version in 2022. FIBO is in use, but the specification is still officially in draft.

FIBO is a development project under the OMG to build a Web Ontology Language (OWL) ontology [OWL, 2012] for describing and sharing financial business data in the form of time series. Time series are the main way data are organized and presented. Data from banks and stock exchanges are the primary sources under consideration. The Dow-Jones Industrial Average is a typical case. Data from statistical agencies were also in scope, so BLS time series are also typical cases.

DDI-CDI is the latest standard in the suite of DDI statistical metadata standards. It is designed to address the needs of describing and integrating data from multiple sources, especially from the statistical point of view. In support of the goal to integrate data from multiple sources, a user needs to be able to describe data in varied structural arrangements. One of these is for multi-dimensional data. The DDI-CDI framework includes the model described here for that purpose.

Both FIBO and DDI-CDI adopted the entire model because the authors of those standards want to be able to describe any multi-dimensional data. The semantic underpinning, flexibility, and completeness of our model were inducements for these adoptions.

Conclusion

This paper describes the efforts at BLS to build a scalable and flexible model for storing and organizing multi-dimensional data. A Measures – People/Places/Things – Time (M-PPT-T) semantic model was described, which accounts for all BLS time series data. The formal UML representation of the model follows directly from the requirements imposed by the M-PPT-T approach. It incorporates some details that are missing from previous work.

Details about universes are given cursory descriptions in other models. They are not seen as having primary importance. Our model shows they are the semantic factor that distinguishes the meaning of

cells in a structural n-cube or distinguishes the meaning of one time series from another. The categories from each dimension are combined into the semantics for the cell, which follows our semantic approach. The combination of unit type and the categories from all the dimensions represents the PPT component.

An additional feature is we do not use the combination of categories as an identifier for the cell. Identifying cells this way depends on the name used for each category, not the combination of meanings of the categories. Since names can change over time and do across programs, identifiers may change. Further, in our model, structural n-cubes are reusable, and once more than one n-cube is constructed from the same structural n-cube these cell identifiers may no longer be unique.

For complex n-cubes, assigning a new variable for each cell is a management and semantic disambiguation nightmare. It is not hard to construct a useful n-cube with thousands of cells, which would require describing and managing thousands of variables with only slight semantic differences among them. Instead, our approach is to specialize the use of a measure so that the semantics of each cell arises from the "intersection" of a universe and a measure (the Qualified Measure in our model). The combined semantics from the components describes the cell rather than using the cell as the point from which descriptions start. The need to manage fewer objects follows.

Finally, as discussed, the BLS the Office of Productivity and Technology implemented the model, and it is the set of requirements for the modernization of the time series database. Two independent standardization efforts adopted the model: The FIBO effort under the OMG and the DDI-CDI effort under the DDI-Alliance. At this writing, the expected release date for DDI-CDI is in early 2023.

References

Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. Commun. ACM 13 (6): 377-387.

Codd, E., S. Codd, and C. Salley. (1993). Providing OLAP to user-analysts: An IT mandate. E. F. Codd and Associates. Vol.32.

DDI-Alliance. (2012) DDI-C Codebook. DDI-C. https://ddialliance.org/Specification/DDI-Codebook/2.5/

DDI-Alliance. (2020a). DDI-L Lifecycle. (2022, Nov). https://ddialliance.org/Specification/DDI-Lifecycle/3.3/.

DDI-Alliance. (2020b). DDI-CDI Cross-Domain Integration. Public Review draft. April 16, 2020). https://ddialliance.org/Specification/ddi-cdi.

OMG. (2017a). Unified Modeling Language. https://www.omg.org/spec/UML/2.5.1/About-UML.

OMG. (2017b) Financial Industry Business Ontology – Indices and Indicator – v1.0 (FIBO). (2022, Nov). https://www.omg.org/spec/EDMC-FIBO/IND/Rafanelli, M., and A. Shoshani (1990) Storm: A statistical object representation model. Proceedings of International Conference on Scientific and Statistical Database Management (SSDBM), pp14-29.

SDMX. (2004). Statistical Data and Metadata eXchange, V1.0 technical specification. (2022, Nov). SDMX. https://sdmx.org/?page_id=18#package

SDMX. (2021). Statistical Data and Metadata eXchange, v3.0 technical specification. (2022, Nov). SDMX. https://sdmx.org/?s=sdmx+3.0+technical+specification

Sundgren, B. (1973). An Infological Approach to Data Bases. Stockholm University and Statistics Sweden, Urval No 7.

UNECE. (2019). Generic Statistical Information Model. (2022, Jan). GSIM. UNECE Supporting Standards Group. https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model.

Van Bracht, E., de Jonge, E. & Kaper, E. (2000) CRISTAL Data Objects. An Object Model for Cubic, Raw, or Intermediate Statistical Data. Netherlands: Statistics Netherlands.

Van Bracht, E. (2004). Cristal - A model for Data and Metadata. Working Paper No. 29, UNECE Work Session on Statistical Metadata (METIS), Geneva. February 2004.

W3C. (1998). Resource Description Framework (RDF). (2022, Nov). https://www.w3.org/RDF.

W3C. (2010). RDF Data Cube Vocabulary. (2022, Nov). https://www.w3.org/TR/vocab-data-cube/.

W3C. (2012). Web Ontology Language (OWL). (2012, Dec). https://www.w3.org/OWL/

Endnotes

[1] Daniel W. Gillman (Gillman.Daniel@BLS.Gov) and Clayton Waring (Waring.Clayton@BLS.Gov); US Bureau of Labor Statistics; 2 Massachusetts Ave, NE; Washington, DC; 20212; USA

[2] Object Management Group - https://www.omg.org/

[3] DDI Alliance – https://ddialliance.org

[4] World Wide Web Consortium - https://www.w3.org/

[5] UML class diagram tutorial - https://creately.com/blog/diagrams/class-diagram-tutorial

[6] Conceptual, logical, and physical models: https://www.guru99.com/data-modelling-conceptual-logical.html

[7] North American Industry Classification System - https://www.census.gov/naics/

[8] Standard Occupational Classification - https://www.bls.gov/soc/

[9] Current Employment Statistics - https://www.bls.gov/ces/

[10] Quarterly Census of Employment and Wages - https://www.bls.gov/cew/