CONFERENCE OF EUROPEAN STATISTICIANS WP.10
Workshop on the Communication of Statistics 20 April 2015
27–29 April 2015, Washington D.C., United States of America

Relational Metadata for Statistical Data APIs (Application Programming interfaces)

Mark Elbert (U.S. Energy Information Administration, United States)

I. SYNOPSIS

Building a robust data Application Program Interface (API) using standards-based metadata fields, along with RESTful interfaces and bulk download files, is the current state of the art for national and international government open data. The U.S. Energy Information Administration (EIA) collects and disseminates statistics, forecasts, and analysis of U.S. and international energy production, consumption, and reserves free of charge to enable policy makers and investors to make informed decisions. This paper describes new data structures that EIA is deploying to convey relational metadata via the agency's RESTful data API. The additional data structures are designed to drive complex interactive data visualizations capable of conveying the facets and complexity of the data. Open-source visualization code that uses the relational metadata is demonstrated and available for download. Relational metadata thus lowers the barrier to visualizing official statistics, satisfying the needs of a wider range of data consumers.

II. INTRODUCTION

The most common organizing principle of data APIs is time, wherein a unique source key is used to request the date-value pairs of a single time series. In some data APIs, such as the U.S. Census Bureau decennial Census API, a unique source key instead returns only a single point in time, but for a set of geographically-defined regions. In this example, these are either the 50 U.S. states or the 3,144 U.S. counties. The organizing principal of these APIs is geography, instead of time.

APIs which return time series are well suited to producing basic time graphs, whereas APIs that publish geographical sets are well suited to producing basic maps. However, the official statistics that our agencies' APIs serve up are usually multidimensional data cubes. For example, EIA publishes the generation of electricity in the U.S. as a four-dimensional data cube, with the dimensions of sector, geography, fuel, and time. Grouping data only along the dimensions of time (time series) or geographical (hereafter referred to as "geosets") fails to convey the other dimensions and the relationships between the various facets in the data cube.

Note that geographical hierarchies within a geoset (e.g. that the state of Virginia is contained within the United States of America) need not be captured in the geoset structure, as geographical relationships are already described in base map definitions.

Ironically, our legacy paper products and their PDF equivalents conveyed relational metadata to the user via table groupings and formatting structures. The illustration below shows a table header from EIA's State Estimation Data System (SEDS) legacy paper-centric report. Note how the vertical columns convey geoset relationships and the horizontal table header conveys the metadata describing the relationships between various geosets. The relationships contained in this table's headers are (1) the total energy consumption to the consumption by fuel source, (2) the total energy consumption to its end-use sectors, and (3) the total consumption of fossil fuels to the consumption by type of fossil fuel.

Table C1. Energy Consumption Overview: Estimates by Energy Source and End-Use Sector, 2012
(Trillion Btu)

		Sources								End-Use Sectors [a]			
		Fossil Fuels											
State	Total Energy [b]	Coal	Natural Gas [c]	Petroleum [d]	Total	Nuclear Electric Power	Renewable Energy [e]	Net Interstate Flow of Electricity [f]	Net Electricity Imports [g]	Residential	Commercial	Industrial [b]	Transportation
Alabama	1,904.7	546.2	682.0	542.7	1,770.8	428.0	247.3	-541.5	0.0	338.1	245.0	839.3	482.3
Alaska	637.3	15.5	347.2	254.3	617.0	0.0	20.3	0.0	(s)	54.6	69.7	323.5	189.5
Arizona	1,407.0	420.6	339.1	489.0	1,248.7	334.6	128.2	-304.6	(s)	385.6	340.7	217.2	463.5
Arkansas	1,064.3	296.2	300.2	320.8	917.2	162.4	115.7	-130.9	0.0	221.7	170.1	391.8	280.6
California	7,640.7	43.8	2,456.3	3,284.0	5,784.1	193.9	805.7	828.6	28.3	1,472.4	1,481.0	1,744.2	2,943.2
Colorado	1,452.4	370.1	456.1	468.7	1,294.9	0.0	106.4	51.1	(s)	336.8	276.6	425.1	413.9
Connecticut	730.3	9.3	236.3	302.3	547.9	179.0	39.8	-36.3	0.0	234.2	183.3	80.0	232.8
Delaware	273.6	17.4	104.4	99.3	221.1	0.0	7.0	45.5	0.0	62.3	56.7	90.2	64.4

Figure 1: Paper-centric report conveying relational metadata

Paper-centric reports have many drawbacks, not the least of which is that in fitting large data cubes onto a two-dimensional printable sheet means only slices of the data cube can be displayed on a single page. In the example above, the time dimension was left out to fit multiple geosets and their relationships onto a single page, and each number displayed is only the latest observation of the cell's time series. Despite these many drawbacks, our legacy paper reports defined geosets and defined relationships between geosets that our more modern APIs usually fail to convey.

III. DATA CUBE STRUCTURE AND RELATIONSHIPS

Figure 2 is an illustration of a simplified demographic data cube demonstrating its structure:

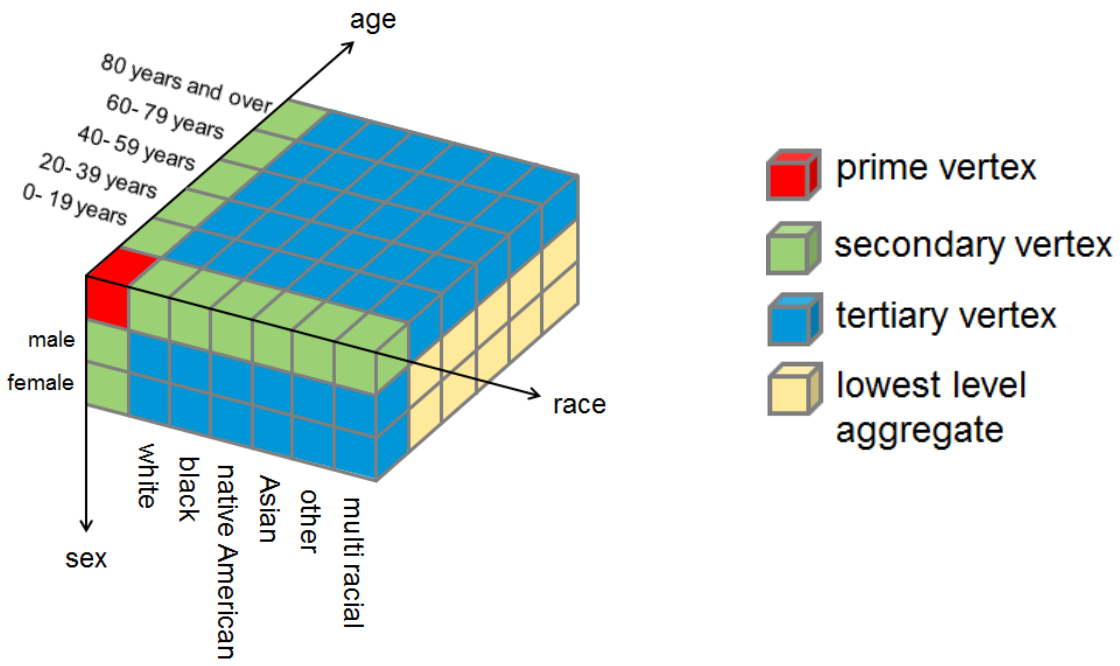

Figure 2: Illustration of a simplified demographic data cube

Each cell in the cube represents geoset (e.g. population of males in [Alaska, Alabama, Arizona, Arkansas…]) and relationships are described at the geoset level. This means we do not have to describe the relations between individual time series or observations, as the geoset relationships are applicable geoset's substructures. Therefore, the geoset concept is indispensable for describing relationships concisely.

Each vertex has one or more relationships to more disaggregated geosets. Sample relationships and potential corresponding visualizations are shown in figures 3 and 4:

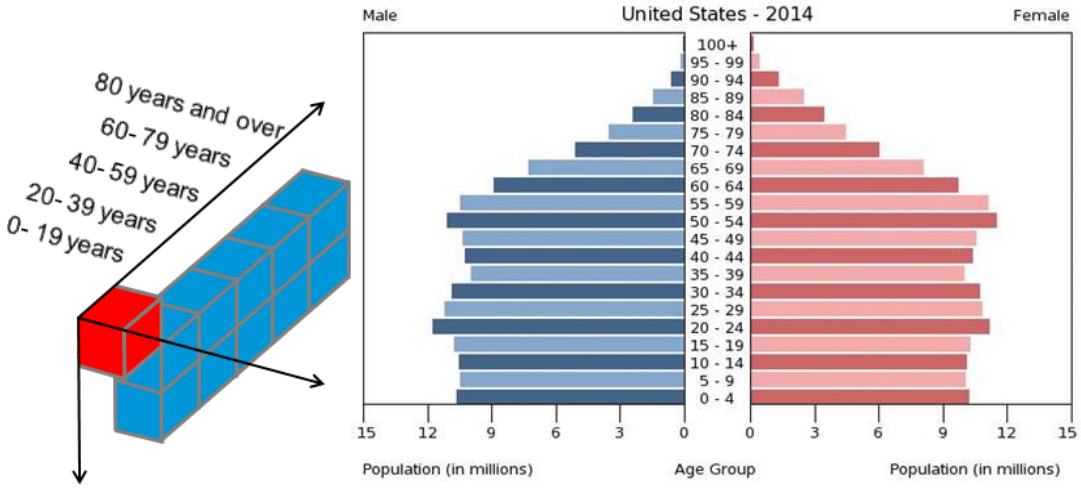

Figure 3: Sample prime-vertex relationship and corresponding visualization

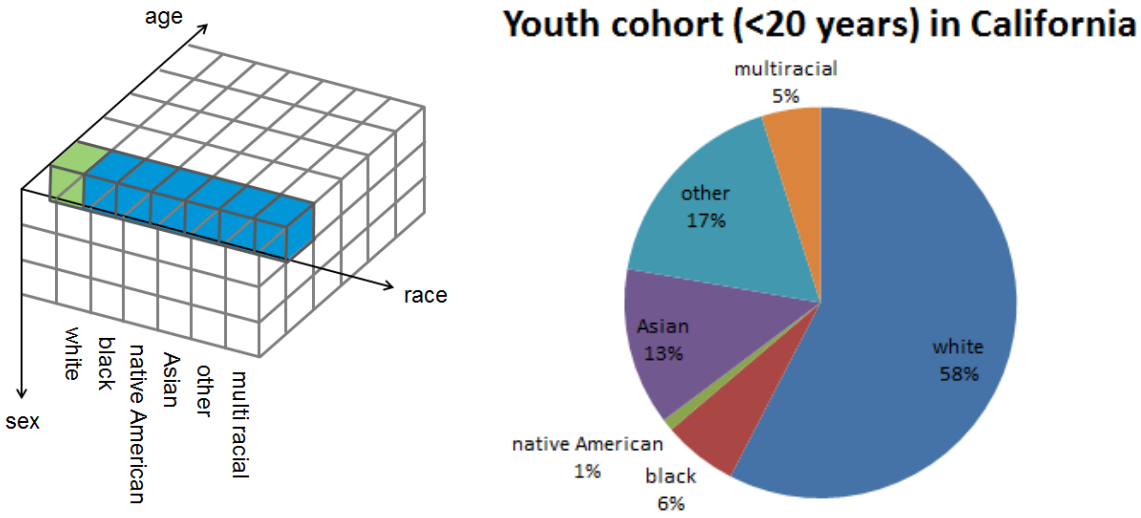

Figure 4: Sample secondary-vertex relationship and corresponding visualization

Each data cube has many such relationships. In our sample demographic data cube, the time series are organized into 126 geosets which have a total of 71 relationships. Putting this into context, such a data cube of the United States might report on 3,190 states, counties and the District of Columbia. So the entire data cube would contain 401,940 time series. The point of this mathematical exercise is to demonstrate that the overhead of adding relational metadata to our existing databases is very manageable.

For the purposes of discussion, this paper used a simplified demographic data cube as population statistics are relatively easy for non-specialist to comprehend. Actual data cube are typically much more complex. Take for example, EIA's publication of net electricity generation data. This data cube has the dimension of sector, fuels, geography and time. The sector dimension (shown below) is hierarchical, which creates additional relationships.

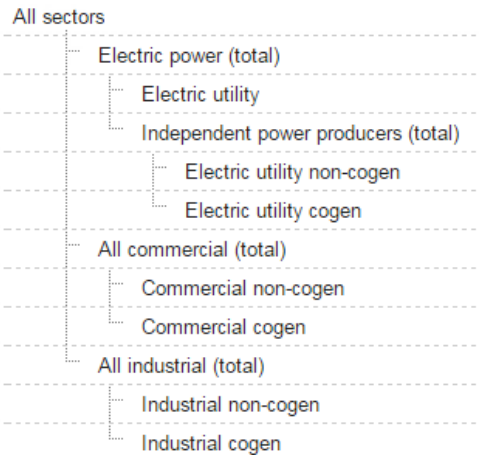

Figure 5: example of a hierarchical data cube dimension

Dimensions whose facets contain hierarchies do complicate the task of capturing all the data relationships. However, capturing this information allows open data APIs capable of driving very interesting data visualizations with dynamic drill-downs and rich data context.

IV. OPEN SOURCE VISUALIZATIONS USING RELATIONAL METADATA FROM THE U.S. ENERGY INFORMATION ADMINISTRATION'S DATA API

EIA has added geoset and relationship structures to its API database. New API commands have been added as well to use the relational metadata to make timely requests for relevant data that can drive advanced data visualizations. The agency has also coded and released an open source JavaScript library to lower the barrier to entry for organizations that want to embed EIA charts and maps on their websites that use relational metadata.

The agency's data API documentation is available on www.eia.gov/beta/api, with links to the visualization libraries on GitHub and usage walk-throughs. Each data series in the API browser provides that series' API call and sample code to embed a chart of the time series, a map of its geoset, and supplementary visualizations as supported by the available relational metadata. In this paper's first illustration, the relationships within energy consumption overview table were described. This relationships have been have been captured and are available via the data API. The API browser's webpages of time series belonging to the "Total Consumption" geoset (e.g. http://www.eia.gov/beta/api/qb.cfm?category=40236&sdid=SEDS.TETCB.US.A) provide links to interactive line charts, maps, pie charts, and clickable map-pit chart interactives. Rather than attempt to explain in words, readers are invited to follow the links see the maps, charts, and advanced data-driven visualization that relational metadata make possible.

V. FINDINGS AND CONCLUSIONS

In EIA's experience, adding relational metadata to the agency's data API has been very manageable, requiring no new systems and a fairly limited amount of labor. By providing open libraries to easily embed a wide array of data visualizations using real time data from the agency, EIA has made itself and its data APIs more relevant to the global discussion of energy production and usage. Supplying data in real time to embedded visualizations makes EIA also a critical service provider to our customers. These new relational metadata structures and their open sourced data visualizations represent a significant step forward in making official statistics accessible, relevant, understandable, and usable to a much wider audience of journalists, businesses, analysts, and lay audiences.