ABSTRACT
The macroeconomic data domain is complex, and the plethora of metadata practices and standards makes using and harmonizing datasets more difficult for researchers. This poster explains the impetus behind and the progress made in the first year of a cooperative project to improve and streamline metadata for economic time series data.

Keywords
Metadata, content standards, data metadata, economic data, time series

INTRODUCTION
The field of economic metadata is fractured and complex, like the data it describes. No one standard or set of standards applies in any area of metadata in this domain: not structure, value, content, or format/technical interchange (Gilliland 2008:3). There are competing schemas used for structuring descriptive data metadata (major schemas are SDMX, DDI, and Dublin Core), and even among producers who use the same schema, there is no accepted implementation standard or completely shared vocabulary. For instance, the SDMX Cross-Domain Code List for geographies recommends ISO 3166 for indicating individual countries, the Nomenclature of Territorial Units for Statistics for coding regions within Europe, and lists but does not explain or standardize “legally defined” regions that do not have codes within ISO 3166 (SDMX 2009:1). Moreover, even this standard’s documentation concedes that these may not be uniformly applied, and that “there is often a need to keep using codes that ISO may suppress or switch them to inactive codes” (SDMX 2009:3).

Other commonly used terminology may hide crucial ambiguities that cause confusion. For instance, the Bureau of Labor Statistics uses six different definitions of the concept of “Boston” (Gillman 2015:14). The common macroeconomic concept of gross domestic product (GDP) can be derived through different mathematical calculations, all of which are expressed as “Gross Domestic Product” (Callen 2012). FRED (http://research.stlouisfed.org/fred2/), a major economic data aggregator, has more than 100 different unit terms representing the U.S. dollar, such as “Current Dollars,” “Dollars,” “USD,” “1982-84 CPI Adjusted Dollars,” “US Dollars, Monthly Level,” etc., across its hundreds of thousands of datasets. This divergence is not particularly surprising; even in the far more unified traditional library world, use of content standards varies significantly (Park and Tosaka 2015: 110).

Although significant work has been done with metadata structures and ontologies to organize the underlying statistical data, including economic statistics such as those from the Bureau of Labor Statistics (Marchionini et al. 2004), neither content standards nor controlled vocabulary/value standards (Gilliland 2008) have been developed for economic time series data such as those found in FRED or viewable in Google’s Public Data Explorer. Time series, defined by the OECD as “a set of regular time-ordered observations of a quantitative characteristic of an individual or collective phenomenon taken at successive, in most cases equidistant, periods/points of time,” (OECD 2005) cannot be properly aligned to one another without congruent metadata describing time, frequency, units, etc. This is particularly difficult to achieve in the case of economic time series in which the data point may be an aggregate or derivative of other data or a ratio of one characteristic to another. A complex unit such as “All employees, 3-month average change, in thousands, seasonally adjusted” (Bureau of Labor Statistics 2015) must be presented clearly and simply to users, but must be broken into its component parts (average, thousands [of persons], seasonal adjustment, etc.) before full interoperability with related data can be achieved.

Comparable projects in other technical fields over the past 20 years have shown the need for a method of expressing technical terms that is both flexible and precise (Cimino 1998; Ma et al. 2010). Issues similar to the description of complex economic variables are found in the biological

ASIST 2015, November 6-10, 2015, St. Louis, MO, USA.

Copyright is retained by the author.
sciences; Courtot et al. (2011) describe markup which defines “neither what the variables and the mathematical expressions represent, nor how they were generated. Where this critical information is communicated through free-text descriptions or non-standard annotations, it can only – if at all – be computationally interpreted with complex text mining procedures” (p. 1) Aggregators of time series data must contend with nearly identical issues.

**PROCESS**

In late September of 2013, in conjunction with a conference for economics and data librarians, the Federal Reserve Bank of St. Louis (informally known as the St. Louis Fed) hosted an informal meeting of representatives of major economic data producers whose data appear in FRED. FRED is a data aggregation website operated by the St. Louis Fed which brings together more than 250,000 time series from nearly 80 sources and provides users tools to compare and combine these data. The represented institutions, which included domestic and international governmental and non-governmental entities, produce data with a wide variety of metadata standards, but have underlying concepts (such as gross domestic product or consumer prices) which span many data series and producers. The aim of the meeting was to determine whether or not formal metadata harmonization across economic data producers was a viable possibility and what form it might take.

Participants agreed that further exploration was warranted and the St. Louis Fed team agreed to take the lead in gathering data from the data producing community. After the initial meeting, the team attempted to draft parameters of a minimal metadata schema (Figure 1) that could be used for time series data and that would allow for automation of data ingest to the FRED system. We then constructed a survey based around the key pieces of metadata that we had identified for FRED use, asking the data producers how they identified and described those aspects of their data in their publicly-available time series.

Although significant differences among data producers were expected, we found no functional areas of content standard overlap between any two respondents to the survey, even those which used a standard structural schema such as SDMX. Unit, geography, title, even time and date were expressed differently by different producers and often even from dataset to dataset. For instance, many respondents indicated that they used ISO codes for countries to indicate geography; however, many also used additional definitions without codes, and more than one used both code and multiple spelled-out versions of a country’s legal name simultaneously. (Note that specific differences by respondent have not been illustrated here; any inclusion of specific survey results in the final poster will depend on permission of the survey respondents.)

**OBSTACLES**

We have already encountered some difficulty in parsing and describing complex units such as “All employees, 3-month average change, in thousands, seasonally adjusted” (Bureau of Labor Statistics 2015). In the graph of such data, it may be opaque to users and computers alike what a data point represents: Is it the difference between two data points (b minus a) or the number of employees? Is the seasonal adjustment applied before or after the calculation of the average? What math underlies the seasonal adjustment?

Despite established metadata best practices that call for mixing and combination of schemas to fully describe resources (Duval et al. 2002), there are no existing schemas or guidelines that are built for economic time series and which are also compatible with a web-ready database structure. Nor are there consistent practices even within the field of economic data or even from a given data producer (Gillman 2015; World Bank). Even a domain-specific controlled vocabulary such as the recently-revised STW Thesaurus for Economics (ZBW 2015) is better suited to library or repository classification use than for granular description of individual data series.

**OUTCOME AND NEXT STEPS**

Based on the survey responses, we determined that agreeing on a standard metadata structure, even a minimal one that could be used alongside other schemas, would be fruitless without establishing rules for standardizing metadata content. The team has decided to turn its attention to creating a standard vocabulary and metadata clearinghouse website that all data producers and consumers can use to trace the differences between datasets and, ideally, use to correct for differences and produce truly harmonized data.
Figure 2. A mockup of a unit entry for US Dollars in the proposed new metadata site

We are working with our data producer partners to create a linked network of terms used in their economic time series metadata in order to make it available to the public and the data community. (See Figure 2)

Our work on standardizing and streamlining units found in FRED has begun, and our planned first steps include creating a style guide for FRED units (“U.S. Dollar,” “National Currency,” etc.) and bringing datasets into compliance with this guide. We hope to have units clarified and defined, and make an initial proof of concept of this metadata dictionary available to our partners, by the end of the year.

Ongoing barriers to progress include conceptual problems such as the appropriate categorization of some data characteristics as unit or unit/title modifier (such as “per capita” or “per hour”) and finding an appropriate software platform to make this information accessible to our partners and the general public.

CONCLUSION

Although many metadata schemas have reached their maturity, content standards and controlled vocabularies in many fields remain un- or under-developed. With a few exceptions (for instance, repositories or sites using the Getty Art and Architecture Thesaurus or Describing Archives: a Content Standard (Gilliland 2008)), implementation of metadata schemas can still vary widely from one application to another. Although subject matter experts in other fields, such as microbiology, have come together to establish content standards for their metadata (Field & Sansone 2006), economics remains a fragmented discipline. We believe that establishing a semi-controlled, heavily documented content vocabulary and metadata dictionary will hopefully lead to a gradual accord of metadata creation practice in the domain of economic time series data.

REFERENCES


<unit>
  <name="thousands of chained 2009 dollars per month" display="Thousands of chained 2009 U.S. Dollars per month"/>
  <type="ratio"/>
  <numerator>
    <characteristics>
      <category="currency" name="U.S. Dollar"/>
    </characteristics>
    <magnitude="1000"/>
    <adjustments>
      <adjustment="true" method.type="chained" method="BEA chained" reference.year="2009"/>
    </adjustments>
  </numerator>
  <denominator>
    <period="month"/>
  </denominator>
</unit>

Figure 3. An early draft markup format


