Data and Metadata from the Terminological Perspective October 2009

By
Daniel W. Gillman[1]
Frank Farance[2]
[1]US Bureau of Labor Statistics, 2 Massachusetts Ave NE, Washington, DC 20212[1]
[2]Farance Inc., 555 Main Street, New York, NY 10044

Abstract
Using the theory of terminology for special languages, we investigate data, their representations, the semantics of data, and allowed computations. These three components for understanding data are related, and terminology supplies the link. The key insight, using terminology, is that a datum is represented by a signifier that stands for a certain kind of concept. The concept is part of the semantics. The kind of concept describes the computation. The paper contains a thorough exploration. A rich descriptive framework is the result, and this framework may significantly alter the way users find statistical data. Currently, users usually must know which agency has the data they are looking for. Reducing or eliminating this dependence will greatly increase the ease and flexibility with which users find US federal statistical data.

Key Words: Concept, computational model, datatype, datum, semantics, terminology, ontology

I. Introduction

This paper contains a discussion of data from the points of view of computation and meaning. Understanding how computation works requires an understanding of datatypes, and understanding how meaning is conveyed, the semantics, requires an understanding of terminology. Computation and meaning are linked, and we describe the linkages in this paper. Consequences for the US federal statistical system are described.

The theory and practice of terminological methods can be used for a better understanding of the meaning of data (data semantics) and for better data exchange among statistical agencies and users (data interoperability), including exchanging the meaning of data consistently (semantic interoperability). The key strategy is to understand a datum as a kind of designation, which is a terminological construct. This understanding was described in our paper, The Nature of Data (Farance and Gillman, 2006).

Computation, semantics, and representation constitute the basic aspects for a description of data. The representational aspect is for saying what the data look like; the computational aspect is for saying how we can compute with the data, e.g., which operations are permissible; and the semantic aspect is about what the data mean. Normally, the computational aspect and semantic aspect are not discussed together. This paper does not so much as lay new ground, it links these aspects to form a unified approach. The unification is achieved through the use of terminological principles. The principles, links, and unification are described.

[1] The opinions expressed in this paper are those of the authors and do not necessarily reflect those of the US Bureau of Labor Statistics

There are potentially significant consequences for the US federal statistical system for this approach. With so many different agencies that produce statistics, users looking for data to answer questions don't always know where to turn. Users don't know and shouldn't need to know which agency has what data. In fact, the users probably haven't even heard of many of them. However, what users want is to be able to find data based on what the data mean, what they can do with it, and what questions they might be able to answer. They want to find data without having to know which agency to get them from, and they want to be able to combine data sets from different sources in order to answer questions the agencies' data don't directly address.

To address these problems, a very sophisticated framework for the semantics and computations for data is needed. The framework addressed here is an attempt to fill this. Typical metadata systems with controlled vocabularies and free text descriptions are not detailed enough. They don't handle concepts well, and concepts underlie much of statistical surveys. They provide the necessary link between what questions are in a user's head and the data that are available to answer them. It is through this linkage that users may achieve these greater aims. Concepts are described through terminological principles, and so this paper is based on those ideas.

II. Background

In terminology, a designation is the association of a concept with a signifier. Signifiers are representational; for instance, they are the strings of alphanumeric characters and other graphs on this page. Designations may be terms, appellations, or symbols. A term, which is a linguistic expression, such as planet, designates a concept that refers to more than one object. An appellation or name, which is also a linguistic expression, such as Neptune, designates a concept that refers to exactly one object. A symbol is any other kind of designation (ISO, 1999). A datum is differentiated from a designation, because a datum is a designation whose concept has a notion of equality defined.

Equality is an essential feature of data, which all data share, for it is necessary to being able to compute with them. Computation may be performed with pencil and paper, an abacus, a slide rule, a calculator, or a computer, but all kinds require copying data from and to memory or storage. Copying data is a basic function of all data processing, and it is basic to the data collected and used in statistical surveys. For example, in survey processing, data are copied from a response form into the computer during key from image; data are transformed and copied back and forth from memory and a database during post-collection processing; and data are copied from the statistical agency server to a user's computer during data dissemination. When a copy is made, we verify it, in theory, by examining it for equality with the original. In practice, this step is sometimes skipped because we know the copying process can be trusted. In fact, the very claim of saying one has a copy implies some kind equality with an original. The ability to determine equality is basic, it enables copying, and copying is required for complex computations and processing, such as in statistics.

It is possible to define equality for some concepts, though we leave alone the question of whether there are concepts for which it cannot. How equality is determined, however, often differs from one concept to another. We use the term value for those concepts that have a notion of equality defined (Farance and Gillman, 2006).

Statisticians also use the word value, and for them it refers to the quantities and categories that statistical data represent. All values from the point of view of statisticians are values from our point of view, because quantities and categories are concepts (as opposed to numerals and codes, which are designations). Notions of equality must have been defined since statistical data are data.

Additionally, values have a notion of repeatability about them. The extension of a concept is the set of all the objects that correspond to that concept, and we always want to be able to determine exactly whether an object is in the extension of a value or not. This reliability is required for accurate classification and is a major determiner of the possibility of measurement error. It turns out, this reliability is impossible to guarantee for any concept (Lakoff, 2002), thus measurement error is inherent to data. However, values exist under the assumption of this repeatability; even though it is impossible to achieve, but with good definitions we can maximize the effect. An example of the reliability problem is with a gender classification. The assumption is that everyone is either male or female, and for most people that is the case; however chromosomal abnormalities and gender identity problems make the distinction very difficult of not impossible to determine sometimes. Other factors, such as masking for disclosure avoidance, can contribute to misclassification, also, but typically they arise as part of the statistical process and are not terminological in nature.

As stated above, the notion of equality defined for one concept may not be the same as that for another. For example, the numeral ’17’ designates the idea of seventeen, the concept corresponding to instances of measures of 17. It is a quantity and a number; and it has the usual equality notion associated with numbers. On the other hand, the letter ‘M’ might designate the idea of being married, a concept corresponding to instances of a non-dissolved marriage. The equality notion here is a little more complex, but it certainly is not the same as for seventeen. We will explore an equality notion for being married later in the paper. The main point is that the notion of equality associated with two concepts may not be the same.

Some values have the same or very similar notions of equality among them, though. For instance, integers have the same notion of equality associated with each one. Sets of such values and their associated signifiers are called value spaces. A value space is one of the three constituents of a datatype, along with a set of assertions and a set of characterizing operations (see section IV) (ISO, 2007a). The kinds of statistical data – nominal, ordinal, interval, and ratio – are sets of datatypes in this sense. The explicit value space is not yet given in each case.

An assertion on a set of values is similar to an axiom. For datatypes, there are five basic kinds of assertions:

- Equality – all datatypes have an assertion about equality, and a notion of equality is defined
- Numeric – some datatypes are based on numeric values, and others not, and this corresponds to the distinction between qualitative and quantitative data in statistics
- Ordering – some sets of values are ordered, such as the integers, and some are not, such as marital status codes
- Exact versus approximate – some values can be fully expressed in a computer, such as codes in a code list, but others cannot, such as irrational numbers

- Boundedness – some sets of values have a bound, such as the natural numbers have a least element, zero, and others sets may have an upper bound, or both; and boundedness refers to an ordering

Assertions are what define datatypes. They are statements that are true about the elements of the value space.

Characterizing operations are the implementations of those assertions that define a datatype. Therefore, because equality is the assertion common to all datatypes, then a notion of equality, i.e., a means for determining equality, must be defined for each datum.

A value space contains a set of values and, therefore, a set of concepts, and a set of concepts structured according to the relations among them is a concept system. So, the set of values in a value space is a concept system, and a concept system with an associated computational model is an ontology. Since the sets of assertions and characterizing operations of a datatype constitute its computational model, i.e., they define and constrain the allowable computations on the values in the value space, then a datatype is an ontology. This will be explored further in the paper.

III. Equality for Values and Value Spaces

A value is a concept with a notion of equality, and a set of values, all with the same notion of equality, and the associated signifiers is a value space. However, how we define equality for concepts may not be immediately clear. In this section, we propose a way to do this for categorical and quantitative data.

Let us start with quantitative data, which are based on numbers. For example, the value seventeen is a number and the concept corresponding to instances of counts of 17. In more mathematical language, this means instances of sets of cardinality 17. Cardinality is defined via set theory and the foundations of mathematics. In fact, the natural (counting) numbers, the integers, the rational numbers, the real numbers, and the complex numbers are each derived from the previous one from set theory, axioms of arithmetic, the notion of limits (from calculus), and roots of equations, in ascending order.

For any concept derived from others, the semantics of the derived concept is the combination of the semantics of the original concepts plus the semantics of the derivation process itself. For instance, real numbers are derived as the limit points of Cauchy sequences of rational numbers. So, one must understand rational numbers and what it means to be the limit point of a Cauchy sequence. Another simpler and more recognizable example of a derived concept is an unemployment rate. Here the base concepts are the labor force and the unemployed. The derivation is to calculate a ratio of these two measures.

Each number system starting from the theory of sets is derived from the others: sets, natural numbers, integers, rational numbers, real numbers, and complex numbers. Therefore, the semantics for a number in each set is determinable from the numbers in the previous (base) set and the derivation from those base numbers to the next set.

Thus, equality of numbers is determined by knowing that the semantics are the same among them, and the question only makes sense if the values being compared come from the same value space. It makes no sense to ask if numbers from different number systems (e.g., integer versus rational) are equal, since the semantics have to be different.

An integer and a rational number cannot be the same since there is an extra derivation for each rational.

For instance, the rational number 17/1 is usually considered equal to the integer 17. And, of course, there are good reasons to say this. However, the numbers are not really the same since different operations exist for the rational number 17/1 than do for the integer 17. We will discuss this more in the next section.

For categorical data, the situation is similar to that for numbers, though it is a little more complex. Equality is still loosely defined in the same way. The semantics of equal values must be the same, and for any comparison to make sense the values must come from the same value space.

How are the semantics of the values defined or derived? We can't rely on the theory of mathematics to ground the semantics for categories the way we do for numbers. Categories, such as gender, occupational, or disease classifications, come from social and cultural conventions. Rather than being contained in mathematical texts, these concepts are defined by statistical agencies or other conventions through consensus. In fact, mathematics is advanced by a kind of consensus, too, through agreement on the correctness of proofs. However, there is much more formality associated with mathematical constructs than those associated with social or cultural conventions.

The semantics of social categories can be found in repositories, registries, or ontologies of concepts managed by statistical offices. The values used in categorical data must refer to these resources for their semantics. They are not as universally agreed upon as the concepts used by mathematicians. This makes finding the semantics for categories more difficult, but the problem of determining equality is conceptually the same as with numbers.

IV. Datatypes and Ontologies

As discussed above, a datatype consists of a value space, a set of assertions, and a set of characterizing operations. A value space is a set of values and their associated signifiers. The assertions are the axioms defining which operations, the characterizing operations, are permitted on the values.

In statistics, the kinds of data – categorical and quantitative – are divided into nominal and ordinal for categorical data and interval and ratio for quantitative data. Nominal, ordinal, interval, and ratio are classes of datatypes, in the sense described above. They are classes of datatypes, rather than datatypes themselves, because the value space is not defined for each. In the following paragraphs, we describe the main assertions and characterizing operations for each class.

Nominal data are the simplest kind to describe as a datatype class. The assertions that define them are

- Equality
- Non-numeric
- Unordered
- Exact
- Unbounded

Marital status codes are an example of nominal data. Note here, numerals may be used as codes, but their order derived from the numbers they usually designate is not allowed.

Ordinal data are like nominal data with the added assertion that they are ordered. By this is meant a linear order, not a partial ordering. The assertions are

- Equality
- Non-numeric
- Ordered
- Exact
- Bounded

The values are non-numeric, even though numerals are often used. If they are, usually the ordering of the numbers the numerals designate is chosen, but ordinal data are not numeric. Generally, there may be several possible orderings for a given value space, and the characterizing operation determines how the ordering is evaluated. A preference scale is an example of ordinal data.

Interval data are quantitative. The assertions are

- Equality
- Numeric
- Ordered
- May be exact or approximate
- May be bounded or not

The characterizing operations define much of the computation, by allowing addition and subtraction, but multiplication and division are not allowed. Temperature in the Fahrenheit or Celsius scale is an example of interval data. For instance, it does not make sense to say 40˚C is twice the temperature of 20˚C. However, it does make sense to say 40˚C is 10˚ warmer than 30˚C. The numbers π and e must always be approximate, and the number ⅛ is exactly represented in decimals: 0.125. So, the exactness assertion depends on the value space, and similarly for the boundedness assertion.

Ratio data are like interval data, they exhibit the same assertions, except they also allow multiplication and division. Because of this, the datatypes in this class are almost always approximate. An example of ratio data is temperature in Kelvin. It is an absolute scale, so multiplication and division may be applied. Now, it does make sense to say 40˚K is twice the temperature of 20˚K (ISO, 2003).

It is well-known, but worth mentioning, that each kind of statistical data has certain statistics that are derivable. Some operations are not allowed, therefore some statistics cannot be produced for some kinds of data. For instance, it does not make any sense to take an average over nominal data, even if numerals are used as the codes for the categories.

Now, we briefly discuss ontologies. The word has meaning in both philosophy and computer science, and here we take the computer science meaning. Ontologies have become much more popular over the last 15 years due to the advent of the Web and more recently the Semantic Web (Berners-Lee, Hendler, and Lassila, 2001). There are about as many definitions of the term as there are researchers in the field. However, we feel that after discussing the concept with said researchers, reading the literature, and observing what ontologies provide in practice, they can be characterized as a concept

system with a computational model defined. This relatively simple definition has some interesting consequences.

As stated before, a computational model for a system consists of a set of assertions, i.e., what the system is allowed to do, and a set of characterizing operations, i.e., how those assertions are calculated. Of course, as discussed before, a datatype, and even a datatype class as defined above, contains a computational model. Therefore, a datatype is an ontology.

Why are ontologies important? Ontologies and the field of formal knowledge representation (Sowa, 2000), e.g., RDF, OWL, and Common Logic (ISO, 2007c), are among the first attempts at a general approach to a formal description of an information system. These formal approaches use first order logic and some variants to try to achieve automated reasoning systems. Statistical metadata systems may be able to take advantage of this new approach. They are designed to describe statistical information systems, e.g., the set of statistical surveys covering labor force in a particular country; and statistics is a fairly well-developed mathematical (thus, formal) framework for designing, manipulating, and analyzing socio-economic data. Thus, the ability to achieve a much more formal and comprehensive approach to metadata is possible. This will be explored further in the next section.

V. Discussion

Up to this point, we have described the following points:

- A datum is a designation of a value
- A value is a concept with a notion of equality
- The proposed notion of equality is a natural appeal to an agreed understanding of values, either quantitative or categorical
- The semantics for a derived concept are due to the semantics of the base concepts and the semantics for any derivations used
- A value space is a set of values
- A datatype is a value space, assertions, and characterizing operations
- Statistical data kinds are classes of datatypes
- A datatype is an ontology
- An ontology is a formal means for organizing data and descriptions

Now, statistics are generated through the application of some function on a set of pre-determined values. The semantics of the statistic, the result, come from the pre-determined values and the semantics of the function itself. An average is the result not only of the averaging function, but also the semantics of the values used in the calculations. So, from the formulas for the statistics and the datatypes that link allowed statistics to each kind of statistical data, we can achieve a formal computational system.

Values are concepts, and they have another interpretation. Values are the properties of a characteristic of a concept, where the concept is really just a population or universe in the normal statistical survey sense. Note, when we use the word characteristic, we do not mean a statistic based on an aggregate. We mean a variable, such as income, applied to each object from the population (Froeschl et al, 2003).

Properties in the terminological sense are determinants, i.e., determined about each object in the extension of a concept. For instance, the specific marital status of a person (single, married, etc) is a determinant. The determinant is a value assigned to a determinable, the characteristic (e.g., marital status) associated with the concept the objects are in the extension for.

We now have the following terminological ideas associated with statistical variables:

- Values are properties of characteristics of concepts
- The characteristics are variables on some population or universe
- The population or universe is the concept whose characteristics are variables
- Properties and characteristics are roles for concepts

This means we can formalize the classifications, variables, and populations into a framework. Computationally, this framework allows one to produce, compare, and combine variables and their classifications automatically for any population and across data sets. It also allows a user to find data based on the meaning of the components of the variables. Thus, this framework is an ontology that formalizes the populations and variables under study.

This produces a link, through the values, between the ontology for computations and the ontology for variables. Together, they produce an ontology for statistical surveys. This represents the potential for a significant shift in the long-term strategy for the functionality of statistical metadata systems. Many authors have discussed these ideas in the past, so this is not new.

There are some significant possibilities here, however, for enhancing the US federal statistical system. There are 14 main statistical agencies in the US and about 60 other agencies with a statistical unit. Each produces its own data under a mandate from the US Congress and the executive branch department the agency falls under. Each disseminates data in its own way from an agency web site, however these web sites are very advanced, using much of the latest technology. Many have data dissemination systems tailored to the specific kinds of data each agency produces, such as the American Fact Finder and Data Web at the US Census Bureau and the One and Multi-Screen Data Search systems at the US Bureau of Labor Statistics. From the site pages themselves, users can select data in several formats, and many kinds of data sets, documentation, and classifications are available.

But, users need to know what web site to go to in advance. Users often don't know and shouldn't need to know which agency has what data. Search engines, such as Google™, help, but they find data such as the current Consumer Price Index or specific classifications such as the Standard Occupational Classification. It is much harder for them to distinguish differences between estimates of total employment. "Why" or "how" questions are particularly difficult unless a specific document was written and placed on the web site. What users want is to be able to find data based on what the data mean, what time period it represents, and what area it covers. How they can compute with it, and what questions they might be able to answer are deeper but just as important.

A few statistical agencies (e.g., ISTAT in Italy) and some other organizations (e.g., US National Cancer Institute and Mayo Clinic) are beginning to use more formalized frameworks for building smarter systems. Though the reasons for building these systems

are different, the efforts are aimed at providing users with more flexibility, and traditional techniques have proved not up to the task.

Unfortunately, the questions the users of statistical data are trying to answer often don't fit neatly into the boxes the agencies have carved out, That is, the data required to answer many questions reside at multiple agency web sites. Currently, there are no standards in place for how an agency should or does disseminate its data and metadata. The agency web sites are not organized in similar ways. Data are presented in several formats, but not every format is available for a particular data set. Worse, metadata for describing the data are mostly not present with the data. Usually, if they are available at all, there are hard to find or search with.

In order for a user to be successful on the web, much of what we have described here must be in place. For a person or a computer to be able to discern more than just basic differences between data sets, a much more sophisticated search and retrieval system must be available. Such a system should be based on the ideas presented here. For there are several advantages to the approach:

- Data may be found without the user needing to know which agency produced it
- Deeper understanding of the data is available on line
- Ability to discern small differences between similar concepts
- Ability to know what kinds of statistics are derivable from each variable
- Ability to answer questions by combining data sets that data from a single agency can't address
- Enhanced ability to harmonize data sets

The road to building such a system within each agency is long. The approach of providing data along the same lines as the business is structured internally may be simple and easy, but it probably doesn't help the users much. So, the steps are
to build a unified system for each agency, thus breaking the adherence to stove pipes
to build a unified system for all the federal statistical agencies, thus providing the user of US federal statistics one view to the data.

VI. Conclusion

This paper contains the outline of a framework for understanding the semantics and computational model for data. The paper shows how the two aspects are linked and how a full description of data is possible from the combination. It concludes with a description of how this framework could be used to provide a single view of US federal statistics. However, the paper does not go into the details for how to implement the framework.

Building a system based on the framework described in the paper will take time and a shared effort. Each agency needs to build its part of the system itself, since it knows its data best. On the other hand, each local system must integrate with all the others. Therefore, the lack of a strong effort to standardize the approach in each agency will result in failure. The paper represents a next step for that standardization effort.

Success will have to be measured incrementally. Many technological choices must be made; some design choices will be difficult to implement; not every program area in each agency will be as enthusiastic a supporter of the effort, especially in the beginning; the

ultimate goal will take a long time to achieve, and management must be on board with the approach from early in its development.

There is also an avenue for further research. The assertions for datatypes are given as choices among five categories. What does the computational model for the semantic side look like? In the paper, we listed several kinds of computations that are necessary, however the details were not given. Is it possible to find assertion types for the semantics as well as the computations? If the answer is yes, then it will be possible to characterize a system for implementing the framework. This in turn easily leads to the necessary standards described just above.

VII. References

Berners-Lee, T., Hendler, J., and Lassila, O. (2001) The Semantic Web. *Scientific American*. May, 2001

CEN (1995). Medical Informatics - *Categorical Structures of Systems of Concepts. Draft*. Brussels: European Committee for Standardization

Farance, F. and Gillman, D. (2006). The Nature of Data. Working Paper #12 in *Proceedings of the UNECE Workshop on Statistical Metadata*, Geneva, Switzerland

Farance, F. and Gillman, D. (2008). Further Developments in the Terminological Principles for Data. Working Paper #12 in *Proceedings of the UNECE Workshop on Statistical Metadata*, Geneva, Switzerland

Froeschl, K., Grossmann, W., & Del Vecchio, V. (2003). The Concept of Statistical Metadata. Deliverable #5 for *MetaNet Project*. Retrieved July 2004 from http://www.epros.ed.ac.uk/metanet/deliverables/D5/IST-1999-29093_D5.doc.

Gillman, D. (2006) Theory and Management of Data Semantics. In D. Schwartz (ed.) *Encyclopædia of Knowledge Management*. Hershey, PA, USA: Idea Group.

Gillman, D. and Johanis, P. (2006). Metadata Standards and Their Support of Data Management Needs. Working Paper #7 in *Proceedings of the UNECE Workshop on Statistical Metadata*. Geneva, Switzerland

ISO (1999). ISO 704: *Principles for terminology*. Geneva: International Organization for Standardization

ISO (2000). ISO 1087-1: *Terminology* – Part 1: *Vocabulary*. Geneva: International Organization for Standardization

ISO (2003). ISO/IEC 20943-3: *Procedures for achieving metadata registry content consisteny*: Part 3 – *Value domains*. Geneva: International Organization for Standardization

ISO (2005). ISO/IEC 11179: *Metadata registries*. Geneva: International Organization for Standardization

ISO (2007a). ISO/IEC 11404: *General purpose datatypes*. Geneva: International Organization for Standardization

ISO (2007b). Draft ISO/IEC 11179-4 (ed 3): *Metadata registries* – Part 4: *Terminological principles for data*. Geneva: International Organization for Standardization

ISO (2007c). ISO/IEC 24706: *Common logic*. Geneva: International Organization for Standardization

Lakoff, G. (2002). *Women, Fire, and Dangerous Things* (Reprint edition). University of Chicago Press

Langefors, B. (1995). *Essays on Infology*. Stockholm: Studentlitteratur

Ogden, C. and Richard, I. (1989). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Harcourt

Sager, J. (1990). *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins

Sowa, J. (2000). *Knowledge Representation*. Brooks Cole Publishing Co., Pacific Grove, CA, 2000