WP. 12
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES
(EUROSTAT)**

**ORGANISATION FOR ECONOMIC
COOPERATION AND DEVELOPMENT
(OECD)
STATISTICS DIRECTORATE**

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Luxembourg, 9-11 April 2008)

Topic 2 (ii), Part B - Metadata Concepts, Standards, Models and Registries

FURTHER DEVELOPMENTS IN THE TERMINOLOGICAL PRINCIPLES FOR DATA [1]

Submitted by United States[2]

I. INTRODUCTION

1. This paper discusses 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. The two are linked, and this paper describes the linkages.

2. 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”, which was part of the proceedings of the METIS 2006 meeting (Farance and Gillman, 2006).

3. In terminology, a designation is the association of a concept with a signifier, such as terms – e.g., planet, whose concepts refer to more than one object – and appellations or names – e.g., Neptune, whose concepts refer to exactly one object – (ISO, 1999). A datum is differentiated from a designation, because a datum is a designation whose concept includes a notion of equality, an essential feature of data.

4. Equality is a feature all data share, and it is essential for being able to compute with them. Computation of all kinds, either on paper, on a hand held calculator or abacus, or on a computer, requires copying data from and to memory or storage, from computer to computer, from paper to computer, from computer to paper or screen, etc. Copying data is a basic function of all data processing, and it is basic to the data collected and used in the statistical survey life-cycle. For example, data are copied from a response

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

[2] Prepared by Frank Farance of Farance Inc and Daniel W. Gillman of the US Bureau of Labor Statistics. Contact frank@farance.com or gillman.daniel@bls.gov.

form into the computer during key from image; data are transformed and copied back and forth from memory and a database during 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 any event, the ability to determine equality is basic, and it is necessary to enable complex computations.

5. It is possible to define equality for some concepts, though how equality is determined often differs from one concept to another. We, the authors, apply the term value to those concepts that include a notion of equality (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), and statistical data are copied just as any data may be, so require a notion of equality. For example, the numeral '17' designates the idea of seventeenness, 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. Further, 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, and we will explore this more later in the paper. The main point is that the notion of equality associated with two concepts may not be the same.

6. 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 are called value spaces, and 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 III) (ISO, 2007a). The kinds of statistical data – nominal, ordinal, interval, and ratio – are datatypes in this sense, without the explicit value space given.

7. A value space is 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, 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.

8. 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 semantical aspect is about what the data mean. Normally, the computational aspect and semantical aspect are not discussed together. This paper does not so much as lay new ground, it links these aspects to form a unified approach. In each of the succeeding sections, we describe equality for values and value spaces; datatypes and ontologies; and statistical data in greater detail. The interrelationships between the parts of the framework are discussed.

II. EQUALITY FOR VALUES AND VALUE SPACES

9. A value is a concept with a notion of equality, and a set of values, all with the same notion of equality, 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. In addition, 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, with good definitions, we can maximize the effect. An example of the reliability problem is with a sex classification. The assumption is that everyone is either male or female, however there are special chromosomal abnormalities that render the distinction moot, and some gender identity problems cause people to want to switch sexes.

10. 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.

11. 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 taking a ratio of these two measures.

12. 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.

13. Thus, equality of numbers is determined by knowing that the semantics are the same among them, and the question only makes sense of 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 cannot be the same since there is an extra derivation for each rational.

14. It should be noted, for instance, that the rational 17/1 is said to be equal to the integer 17. And, of course, there are many more examples like 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.

15. 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 the values must be the same, and for any comparison to make sense the values must come from the same value space.

16. How are the semantics of the values defined or derived? We can’t rely on the theory of mathematics to ground the semantics the way we do for numbers. Categories, such as from sex, industrial, 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, but it doesn’t feel that way, nor is it described that way by practitioners.

17. 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 point 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.

III. DATATYPES AND ONTOLOGIES

18. A datatype consists of a value space, a set of assertions, and a set of characterizing operations. A value space, as previously discussed, is a set of values. The assertions are the axioms defining which operations, the characterizing operations, are permitted on the values.

19. 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.

20. Nominal data are the simplest kind to describe as a datatype class. There are no assertions except equality, exactness, non-numeric, and cardinality. This means it is possible to determine equality on the values; each value is exactly represented in a computer, as opposed to being an approximation, as in some real numbers; the values are non-numeric; and it is possible to know the cardinality of a set of nominal data. Typical characterizing operations will be the means to determine equality; and the means to know the cardinality of the value space. An example of nominal data is a sex classification.

21. 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 values may also be numeric, as long as performing arithmetic is not part of the assumptions. There may be several possible orderings for a given value space, and the characterizing operation determines how the ordering is evaluated. An example of ordinal data is a set of preferences.

22. Interval data are quantitative. They have equality, of course, are ordered, can be exact or approximate, are numeric, may be bounded though usually not, and have cardinality associated. The characterizing operations define much of the limitations, 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.

23. Ratio data are like interval data, except they also allow multiplication and division. Because of this, the datatypes in this class are usually 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).

24. 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.

25. 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. 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. This relatively simple definition has some interesting consequences.

26. 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 carried out. 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.

27. 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.

IV. DISCUSSION

27. 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 some concept are due to the semantics of the basic concepts assumed 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

28. 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.

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

30. Properties in the terminological sense are determinants, i.e., determined about each object in the extension of a concept. For instance, the specific color of the iris of a person's eyes is a determinant. The determinant is a value assigned to a determinable, the characteristic (e.g., eye color) associated with the concept the objects are in the extension for.

31. Now, we have the following terminological ideas associated with statistical variables:

- Values are properties of characteristics of concepts
- The characteristics are variables on some population
- The population 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 variables and their classifications automatically for any population. Thus, this framework is an ontology that formalizes the populations and variables under study.

32. 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. We hope this paper provides more of a roadmap than has been achieved previously.

IV. REFERENCES

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 presented at 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.) Encyclopaedia 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 presented at 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: Procedured 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